Thesis for the Degree of Licentiate of Philosophy

Nonparametric Linkage Power by Simulations and Multivariate Normal Approximations

Mikael Knutsson

Department of Mathematical Statistics
Chalmers University of Technology and Göteborg University
Göteborg, Sweden 2000

Nonparametric Linkage Power by Simulations and Multivariate Normal Approximations
Mikael Knutsson

ISSN 0347-2809/NO 2000:10
Department of Mathematical Statistics
Chalmers University of Technology and Göteborg University
SE-412 96 Göteborg
Sweden
Telephone +46-(0)31-772 1000

Chalmers University of Technology
Göteborg, Sweden 2000

Abstract

A simulation method for generation of marker data given a specified genetic model is presented, especially for nuclear families with affected sib-pairs. We describe how powers in nonparametric linkage analysis can be obtained, combining the marker simulations with calculations by the GENEHUNTER software and a multivariate normal approximation. This stepwise method is illustrated and tested by an example, using MS data. Finally, the approach is used to find the appropriate thresholds in partial genome scans.

Keywords: Affected sib-pair, GENEHUNTER, Multiple Sclerosis, the multivariate normal distribution, NPL-scores

Acknowledgements

To my supervisor Professor Olle Nerman, Marita Olsson, Staffan Nilsson, Ulrica Olofsson (formerly known as roommate), Oluf Andersen and Sara Haghighi at Sahlgrenska University Hospital, and to everyone who has encouraged or in some weird sense inspired me along the road. Thank You!

Mikael Knutsson,
Göteborg, February 2000

Contents

Chapter 1

Introduction

In the search for (disease causing) genes, the use of statistical analysis is a necessity. Undertaking a whole genome scan, blood is collected from a large number of individuals, ascertained through some specified sampling criteria. The scanning process searches through the *genome* (the genetic material in the chromosomes) and types the individuals at a large number of known positions on the chromosomes called *markers*, which are segments of DNA whose pattern of inheritance can be determined (the "typing" can be performed, using *electrophoresis*). Having typed individuals, *linkage analysis* can be used to find the relative position of the gene. Since the cosegregation between two locations is dependent of the distance between them, the cosegregation between a disease causing gene and the markers in its neighbourhood will be high. The crucial idea of linkage analysis is that if affected relatives share a "significant" amount of similarities at a certain marker locus, the conclusion is that there is linkage between this marker and a disease related gene, i.e. there is a disease gene located in the neighbourhood of the marker (at least on the same chromosome). If the scan suggests some candidate regions for disease genes, it can be followed up by further studies, using a denser map of markers in the particular regions of interest. Obviously, there are problems involved in the process concerning the multiple testing and it is hard to get clear-cut results, especially for complex diseases where several genes (+ environment) interacts, each one providing only a small contribution to the disease so that the tendency of cosegregation is not very strong at any marker loci. For more details, the reader is encouraged to check the existing literature on the subject ([Ott91], [GMS+96]).

The main purpose of this thesis is to find a method to estimate the power of a future linkage study. The method was brought to life during the planning stage of an affected sib-pair study, regarding the *Multiple Sclerosis* disease.

Before starting to collect real data (from blood samples etc), it is important to get an idea of how much and what kind of data (number of families, family sizes, structures etc) that will be needed to, say, locate a disease susceptibility locus if there are any or rather to detect linkage to such a locus. Of course, there is no way to find an exact answer to this question, due to the limited amount of knowledge regarding the disease and the many assumptions that have to be made to make calculations feasible, but at least some upper bounds for the power can be obtained, providing answers to questions like:

- "We will be able to collect data from 50 families of a certain structure, will it be enough to have a fair chance of finding linkage to a disease locus, when in fact there is such a locus?"
- "We have collected blood from two affected sibs. There is a third healthy (unaffected) sib. How much better results can be obtained from collecting blood also from this third sib? Is it just a waste of time and money?"
- Before undertaking a large-scale study we wish to be sure to find "something" with probability at least 0.8. What is the implied lower bound of the number of families to collect?"
- "To confirm whether an individual has a certain disease, the individual has to undergo a trying procedure, associated with some risks. Sibs of known affected patients have already been tested. Is it worth the effort of testing the parents of the patients?"

The data will consist of *marker* genotypes (from blood samples) and disease status (affected/unaffected, in our situation from lumbal puncture). The natural first step in the power calculations is therefore to, given a specified genetic model (Chapter 2), simulate marker data for a large number of families (Chapter 3).

Having a large amount of marker genotypes the appropriate linkage statistic can for each family and at every marker locus be calculated. Here the focus will be on the NPL (nonparametric linkage) score, implemented in the GENEHUNTER software [KDRDL96] [KL98] and the score distribution is approximated by a multivariate normal distribution (Chapter 4).

In Chapter 5, the method is illustrated and tested by an example, using the MS data and in Chapter 6 the same approach is used to find the appropriate thresholds and pointwise levels of significance to use, for instance when performing partial genome scans. Unlike the power simulations, the simulations used to find thresholds are performed under the null hypothesis of no linkage. Finally, some concluding remarks (Chapter 7).

Chapter 2

The Genetic Model

In the model we assume a single disease locus $\mathcal{D}$ (location of the disease related gene), with a trait (disease) allele A and a normal allele a (or rather a set of normal alleles, i.e. $a = \{a_1, a_2, \ldots, a_d\}$), existing in the population of interest in the proportions p_A and p_a (of course, $p_a = 1 - p_A$). Each individual has a set of two alleles at this disease locus, belonging to different chromosomes, one inherited from the mother and the other from the father. The possible sets of two alleles give rise to the "three" *genotypes* AA, Aa and aa, existing in the population under study in the proportions p_A^2, $2p_Ap_a$ and p_a^2, under the assumption of *Hardy-Weinberg equilibrium.* The number three put between quotation marks because Aa being the set

$$\{A \cap (\cup_{i=1}^d a_i)\}$$

of all genotypes with exactly one trait allele and

$$aa = \{(\cup_{i=1}^d a_i) \cap (\cup_{i=1}^d a_i)\}$$

all genotypes without the trait allele. Other "standard assumptions" which are made are crossovers according to Haldane's [Hal19] map function (see Chapter 3) and *random mating*, i.e. two individuals (male and female) mate independently of each others genotypes. Of course, it is not clear whether this (random mating) is a realistic assumption. However, it is hard to substitute and makes calculations possible (moreover, it is also a natural framework in which the Hardy-Weinberg assumption is plausible).

The *phenotype* (affected/unaffected) of an individual is conditioned on the genotype, assumed independent of the genotypes and phenotypes of all the other individuals in the family or pedigree. The penetrance vector

$$(f_{AA}, f_{Aa}, f_{aa})$$

is defined as the probabilities of being affected conditioned on the genotypes, i.e.

$$f_G = P(\text{being affected} \,|\, \text{genotype } G),$$

for the genotypes $G = AA, Aa, aa$. The penetrance vector describes the affection mechanism;

- $f_{AA}, f_{Aa} \neq 0$, $f_{aa} = 0$: *dominant* model, i.e. at least one trait allele is needed to get affected (if $f_{AA} = f_{Aa} = 1$, we say we have a *fully penetrant* dominant model)

- $f_{AA} \neq 0$ and $f_{Aa} = f_{aa} = 0$: *recessive* model, where two trait alleles are needed for affection ($f_{AA} = 1$ gives a fully penetrant recessive model)

- $f_{aa} \neq 0$: allows for *phenocopies*, i.e. there is a small probability of being affected through environmental causes, without having the trait allele. This might be a different disease, which is impossible to distinguish from the one under study.

The simplified assumption of f_{AA}, f_{Aa} and f_{aa} being constants is made, which in many diseases is an unrealistic assumption because the affection event often depends on, say, the individual's age. This is the case in for instance the MS disease, where the affection usually occurs in the ages 20-40. Therefore it seems reasonable to use some kind of age dependent survival functions to model the penetrance events. Such functions would probably be easy to incorporate into the simulation method (however not implemented) and thus the penetrances are treated as constants with no loss of generality.

From p_A, f_{AA}, f_{Aa} and f_{aa} the population prevalence

K_P - the proportion of affected individuals

and the sib prevalence

K_S - the probability of being affected for
an individual with an affected sib

can be calculated. One obtain (in this model)

$$K_P = f_{AA}p_A^2 + 2f_{Aa}p_Ap_a + f_{aa}p_a^2$$

and

$$
\begin{aligned}
K_S &= \frac{\mathsf{P}(\text{2 affected offspring in a two-sib family})}{K_P} \\
&= {K_P}^{-1} \sum_{G_M} \sum_{G_F} \mathsf{P}(ASP \mid G_M, G_F) \mathsf{P}(G_M, G_F) \\
&= {K_P}^{-1} [p_A^4 f_{AA}^2 + 4p_A^3 p_a (\frac{1}{4} f_{AA}^2 + \frac{1}{4} f_{Aa}^2 + \frac{1}{2} f_{AA} f_{Aa}) + 2p_A^2 p_a^2 f_{Aa}^2 \\
&\quad + 4p_A^2 p_a^2 (\frac{1}{16} f_{AA}^2 + \frac{1}{16} f_{aa}^2 + \frac{1}{4} f_{Aa}^2 + \frac{1}{4} f_{AA} f_{Aa} + \frac{1}{4} f_{Aa} f_{aa} + \frac{1}{8} f_{AA} f_{aa}) \\
&\quad + 4p_A p_a^3 (\frac{1}{4} f_{Aa}^2 + \frac{1}{4} f_{aa}^2 + \frac{1}{2} f_{Aa} f_{aa}) + p_a^4 f_{aa}^2],
\end{aligned}
$$

where G_M and G_F denote the genotypes of the mother and father and ASP denotes the event that the parents and the two offspring constitute an affected sib-pair family, i.e. a family with two affected offspring (identical twins excluded). The six terms in the K_S expression are associated with parental genotypes (AA, AA), (AA, Aa), (AA, aa), (Aa, Aa), (Aa, aa) and (aa, aa), where for instance (AA, Aa) is the notation for the event

$$\{(G_M = AA \cap G_F = Aa) \cup (G_M = Aa \cap G_F = AA)\}.$$

However, the values of p_A, f_{AA}, f_{Aa} and f_{aa} are usually not known, but some sample based estimates of the prevalences K_P and K_S can be used to estimate the parameters using some "backwards calculations". In the recessive case there are only two parameters to estimate (p_A and f_{AA}) and these can be unambiguously estimated through the equations

$$\begin{cases} K_P = p_A^2 f_{AA} \\ K_S = \frac{f_{AA}}{4}(1 + p_A)^2 \end{cases},$$

leading to the expression

$$\lambda_S = \frac{(1 + p_A)^2}{4p_A^2},$$

where

$$\lambda_S = \frac{K_S}{K_P}$$

is the risk ratio for sibs of affected individuals compared to the population in general. In the dominant case however, there are three parameters to estimate (p_A, f_{AA} and f_{Aa}) and thus one of the parameters has to be fixed to obtain the others. By letting $f_{AA} = f_{Aa} = f$ the prevalence equations become of the form

$$\begin{cases} K_P = f p_A(2 - p_A) \\ K_S = \frac{f}{4(2-p_A)}(p_A^3 - 6p_A^2 + 5p_A + 4) \end{cases},$$

resulting in

$$\lambda_S = \frac{p_A^3 - 6p_A^2 + 5p_A + 4}{4p_A(2 - p_A)^2}.$$

In the model we also need a set of marker loci, flanking the disease locus (on the same chromosome), and the "distances" in centiMorgan (cM) between these markers and the disease locus (the human genome has a total "length" about 33M=3300cM). The distance in Morgan between two loci is defined as the expected number of crossovers between the loci (expectation of one crossover per Morgan, see Chapter 3 for more details). Let N be the the number of markers to be generated and denote those markers by

$$\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_N.$$

Let n_i, $i = 1, 2, \ldots N$, be the number of possible alleles at $\mathcal{M}_i$ and $\{m_{i1}, m_{i2}, \ldots, m_{in_i}\}$ that allele set. Corresponding to the ith allele set there is a frequency vector

$$\boldsymbol{q}_{\mathcal{M}_i} = (q_{m_{i1}}, q_{m_{i2}}, \ldots, q_{m_{in_i}}),$$

where $q_{m_{ij}}$ is the population frequency of allele m_{ij}, satisfying the equality

$$\sum_{j=1}^{n_i} q_{m_{ij}} = 1, \quad i = 1, 2, \ldots, N.$$

Of course, the assumption of random mating holds for the marker genotypes as well and *linkage equilibrium* is assumed, i.e.

$$\mathsf{P}(\mathcal{D} = d, \mathcal{M}_1 = m_{1j_1}, \mathcal{M}_2 = m_{2j_2}, \ldots, \mathcal{M}_N = m_{Nj_N}) = p_d \prod_{i=1}^{N} q_{m_{ij_i}}.$$

Suppose the disease locus is located between the Kth and the $(K+1)$th marker locus and let $\boldsymbol{C}$ be a distance vector

$$\boldsymbol{C} = (c_1, c_2, \ldots, c_N),$$

where

$$
\begin{array}{ll}
c_1 & \text{denotes the distance in cM between } \mathcal{M}_1 \text{ and } \mathcal{M}_2 \\
c_2 & \text{between } \mathcal{M}_2 \text{ and } \mathcal{M}_3 \\
\vdots & \\
c_K & \text{between } \mathcal{M}_K \text{ and } \mathcal{D} \\
c_{K+1} & \text{between } \mathcal{D} \text{ and } \mathcal{M}_{K+1} \\
\vdots & \\
c_N & \text{between } \mathcal{M}_{N-1} \text{ and } \mathcal{M}_N.
\end{array}
\quad .
$$

From these distances recombination probabilities can be calculated, using Haldane [Hal19] or some other map function (see Chapter 3). In the model presented here, there are no differences in recombinations between the sexes, i.e. recombinations between marker loci on chromosomes of males occur with the same probabilities as for females. However, such differences can easily be incorporated by using different distance vectors for the sexes (substitute $\boldsymbol{C}$ with some vectors $\boldsymbol{C}_m$ and $\boldsymbol{C}_r$).

Chapter 3

Simulation of Marker Data

Here, the focus will be on ASP (affected sib-pair) families with two parents and two affected offspring, motivated by their usefulness (and usualness) in medical studies. One obvious way to generate marker data for such families is of course to randomly assign alleles at the disease and marker loci to each one of the parents (according to the population frequencies), generate genotypes and affection status of two children and if the children are affected, keep the family, else throw it away and start all over with fresh parents. This must be the easiest way to generate the families, but also very time consuming. Instead, the strategy will be to first consider the assumed disease related locus $\mathcal{D}$ and obtain genotypes at this locus for ASP families before considering any markers. This is done by simulating directly from the conditional distribution of family genotypes, given the information that the two sibs are affected. I.e. sampling from the conditional distribution

$$\mathsf{P}(G_M, G_F, G_{O_1}, G_{O_2} \,|\, ASP),$$

where G_M, G_F are the genotypes at locus $\mathcal{D}$ of the mother and father and G_{O_i} the genotype of the ith offspring, rather than to use a rejection procedure (i.e. generate families until the "the right ones" are obtained) which tends to be very time consuming for low prevalences and penetrances and therefore inefficient when dealing with rare diseases. By using this method the computation time will be independent of the prevalences and penetrances (equal computation times for all diseases). To gain more in time we actually simulate parental genotypes from the distribution

$$\mathsf{P}(G_M, G_F \,|\, ASP)$$

first and then, after having obtained G_M and G_F, simulate offspring from

$$\mathsf{P}(G_{O_1}, G_{O_2} \,|\, G_M, G_F, ASP).$$

Parental group PG_i	Pop Freq $\mathsf{P}(PG_i)$	Possible offspring[1] OG_{ij}	$\mathsf{P}(ASP \mid OG_{ij})$
(AA, AA)	p_A^4	(AA, AA) w.p. 1	f_{AA}^2
(AA, Aa)	$4p_A^3 p_a$	(AA, AA) w.p. $\frac{1}{4}$ (AA, Aa) w.p. $\frac{1}{2}$ (Aa, Aa) w.p. $\frac{1}{4}$	f_{AA}^2 $f_{AA} f_{Aa}$ f_{Aa}^2
(AA, aa)	$2p_A^2 p_a^2$	(Aa, Aa) w.p. 1	f_{Aa}^2
(Aa, Aa)	$4p_A^2 p_a^2$	(AA, AA) w.p. $\frac{1}{16}$ (AA, Aa) w.p. $\frac{1}{4}$ (AA, aa) w.p. $\frac{1}{8}$ (Aa, Aa) w.p. $\frac{1}{4}$ (Aa, aa) w.p. $\frac{1}{4}$ (aa, aa) w.p. $\frac{1}{16}$	f_{AA}^2 $f_{AA} f_{Aa}$ $f_{AA} f_{aa}$ f_{Aa}^2 $f_{Aa} f_{aa}$ f_{aa}^2
(Aa, aa)	$4p_A p_a^3$	(Aa, Aa) w.p. $\frac{1}{4}$ (Aa, aa) w.p. $\frac{1}{2}$ (aa, aa) w.p. $\frac{1}{4}$	f_{Aa}^2 $f_{Aa} f_{aa}$ f_{aa}^2
(aa, aa)	${p_a}^4$	(aa, aa) w.p. 1	f_{aa}^2

Table 3.1: Inheritance probabilities.

The correct distributions can be obtained from Table 3.1. In the first column, the table displays the six possible parental groups, say

$$PG_1, PG_2, \ldots, PG_6.$$

Each parental group occurs with a probability

$$\mathsf{P}(PG_i), \quad i = 1, 2, \ldots, 6$$

(column 2). Given the parental group PG_i, there are some number $n(PG_i)$ of possible offspring groups, say

$$OG_{i1}, OG_{i2}, \ldots, OG_{in(PG_i)},$$

each with probability

$$\mathsf{P}(OG_{ij} \mid PG_i), \quad j = 1, 2, \ldots, n(PG_i),$$

conditioned on the parental group (column 3). The fourth column shows the conditional probability of obtaining an ASP family (both offspring affected)

[1] w.p. - with probability.

for the different offspring groups. The conditional probability for PG_i, conditioned on the family being an ASP family, is

$$\begin{aligned}
\mathsf{P}(PG_i \mid ASP) &= \frac{\mathsf{P}(PG_i \cap ASP)}{\mathsf{P}(ASP)} \\
&= \frac{\sum_{j=1}^{n(PG_i)} \mathsf{P}(OG_{ij} \cap PG_i \cap ASP)}{\mathsf{P}(ASP)} \\
&= \frac{\sum_{j=1}^{n(PG_i)} \mathsf{P}(OG_{ij} \cap PG_i)\mathsf{P}(ASP \mid OG_{ij} \cap PG_i)}{\mathsf{P}(ASP)} \\
&= \frac{\sum_{j=1}^{n(PG_i)} \mathsf{P}(PG_i)\mathsf{P}(OG_{ij} \mid PG_i)\mathsf{P}(ASP \mid OG_{ij})}{\mathsf{P}(ASP)} \\
&= \frac{\mathsf{P}(PG_i) \sum_{j=1}^{n(PG_i)} \mathsf{P}(OG_{ij} \mid PG_i)\mathsf{P}(ASP \mid OG_{ij})}{\mathsf{P}(ASP)},
\end{aligned}$$

where

$$\mathsf{P}(ASP) = \sum_{i=1}^{6} \mathsf{P}(PG_i) \sum_{j=1}^{n(PG_i)} \mathsf{P}(OG_{ij} \mid PG_i)\mathsf{P}(ASP \mid OG_{ij})$$

($\mathsf{P}(ASP \mid OG_{ij} \cap PG_i) = \mathsf{P}(ASP \mid OG_{ij})$ since the probability for the two offspring to be affected does not (by assumption) depend on the parental genotypes when conditioning on the offspring genotypes). For a parental group (A_1A_2, A_3A_4), the parental genotypes becomes

$$\begin{cases} (G_M = A_1A_2, G_F = A_3A_4) & \text{w.p. } \frac{1}{2} \\ (G_M = A_3A_4, G_F = A_1A_2) & \text{w.p. } \frac{1}{2} \end{cases}.$$

Conditioning on PG_i and the family being ASP, the probability for offspring group OG_{ij} is

$$\begin{aligned}
&\mathsf{P}(OG_{ij} \mid PG_i, ASP) \\
&\quad = \frac{\mathsf{P}(OG_{ij} \cap PG_i \cap ASP)}{\mathsf{P}(PG_i \cap ASP)} \\
&\quad = \frac{\mathsf{P}(OG_{ij} \cap PG_i \cap ASP)}{\sum_{j^*=1}^{n(PG_i)} \mathsf{P}(OG_{ij^*} \cap PG_i \cap ASP)} \\
&\quad = \frac{\mathsf{P}(OG_{ij} \mid PG_i)\mathsf{P}(ASP \mid OG_{ij})}{\sum_{j^*=1}^{n(PG_i)} \mathsf{P}(OG_{ij^*} \mid PG_i)\mathsf{P}(ASP \mid OG_{ij^*})}.
\end{aligned}$$

Having obtained the offspring group, G_{O_1} and G_{O_2} are picked at random (we also need to keep track of which chromosomes the alleles are inherited from). For instance, one parent becomes homozygous AA and the other heterozygous Aa with probability

$$\frac{4p_A^3 p_a(\frac{1}{4}f_{AA}^2 + \frac{1}{2}f_{AA}f_{Aa} + \frac{1}{4}f_{Aa}^2)}{\mathsf{P}(ASP)} = \frac{p_A^3 p_a(f_{AA} + f_{Aa})^2}{\mathsf{P}(ASP)}$$

and given these parental genotypes the two offspring are assigned genotypes

$$\begin{cases} (AA, AA) & \text{with probability } \frac{f_{AA}^2}{(f_{AA}+f_{Aa})^2} \\ (AA, Aa) & \text{with probability } \frac{2f_{AA}f_{Aa}}{(f_{AA}+f_{Aa})^2} \\ (Aa, Aa) & \text{with probability } \frac{f_{Aa}^2}{(f_{AA}+f_{Aa})^2} \end{cases}.$$

With probability

$$\frac{2p_A^2 p_a^2 f_{Aa}^2}{\mathsf{P}(ASP)}$$

both parents will instead become homozygous, one with genotype AA and the other aa. Given these parental types the offspring will be heterozygous at $\mathcal{D}$.

That the sampling scheme gives the desired distribution follows from the equality

$$\begin{aligned} &\mathsf{P}(G_M, G_F, G_{O_1}, G_{O_2} \mid ASP) \\ &\quad = \mathsf{P}(G_M, G_F \mid ASP)\mathsf{P}(G_{O_1}, G_{O_2} \mid G_M, G_F, ASP). \end{aligned}$$

It is easy to see that the formulas for $\mathsf{P}(PG_i \mid ASP)$ and $\mathsf{P}(OG_{ij} \mid PG_i \cap ASP)$ still hold when adding more offspring, simply by modifying the ASP condition to denote at least two offspring affected, exactly two offspring affected, at least three offspring affected and so on. However, the number of possible offspring groups will increase drastically and a larger table will be needed.

This first simulation step results in the genotypes at $\mathcal{D}$ for the families and also information about the inheritance (from parents to offspring). If the parental affection status are of interest, these are easily checked. For each parent, simply generate a random number U from the $Uniform(0, 1)$ distribution and compare it to the penetrance f_G, associated with the parental genotype G.

$$\begin{cases} U < f_G & \Rightarrow \text{ affected} \\ U \geq f_G & \Rightarrow \text{ unaffected} \end{cases}.$$

For each family the next step is to randomly assign marker alleles for each of the parental chromosomes according to the population frequencies $\{q_{m_{ij}}\}$.

The final step is then to pass down marker alleles from the parents to the offspring, but first let us try to understand the underlying processes in the *meioses*, motivating the use of Haldane's map function [Hal19].

The meioses is the cell division process where the chromosomes, later passed down to the offspring, are constructed. In the meioses two cell divisions occur, resulting in four cells, each containing one chromosome from each of the pairs in the original cell (see [GMS+96] for more details). Consider two copies of each of the two chromosomes of a parental pair (see Figure 3.1(a)). The non-sister chromosomes start to cross over each other (see Figure 3.1(b)). Two point processes are needed to describe the crossovers [SQ99]. The first process determines the locations of the crossovers and for this a Poisson point process can be chosen. The second process tells us between which chromosomes the crossovers occur and the reasonable choice for this process is to let each of the four pairs of non-sister chromosomes be the pair involved in the crossover with probability 1/4. The assumption of *no chromatid interference* is made, i.e. the pair involved in a crossover is independent of all the previous crossovers and the location of the cross-over.

Now, consider two marker loci $\mathcal{M}_1$ and $\mathcal{M}_2$ at distance c_1 from each other. What is the probability of recombination between these loci? Let $N_{1,2}$ be the number of crossovers of the four non-sister chromosomes occurring between $\mathcal{M}_1$ and $\mathcal{M}_2$. The result of the meisoses is four possible chromosomes. If $N_{1,2} = 0$, none of the four choices will be recombinant and the probability of recombination between $\mathcal{M}_1$ and $\mathcal{M}_2$ will be 0. If $N_{1,2} = 1$ (like in Figure 3.1(b)), the meioses results in two recombinant and two non-recombinant chromosomes, i.e. recombination with probability 1/2. As it turns out, 1/2 will be the probability for all $N_{1,2} > 0$, i.e. the probability of recombination between $\mathcal{M}_1$ and $\mathcal{M}_2$ is

$$\frac{1}{2}\mathsf{P}(N_{1,2} > 0).$$

If $N_{1,2} \sim Poisson(2c_1/100)$, which follows from the assumption of the Poisson point process, Haldane's map function

$$\frac{1}{2}[1 - \exp(-2c_1/100)]$$

is obtained. Although the Poisson assumption is not the only the way to obtain this map function, it is the only natural way to get it.

In principle, given the parental marker pattern, the markers passed down from one parent to one of the offspring are simulated as two discrete inhomogeneous Markov chains, to the left and to the right of the disease allele, each with two possible states, say 0 and 1, where state 0 denotes the chromosome

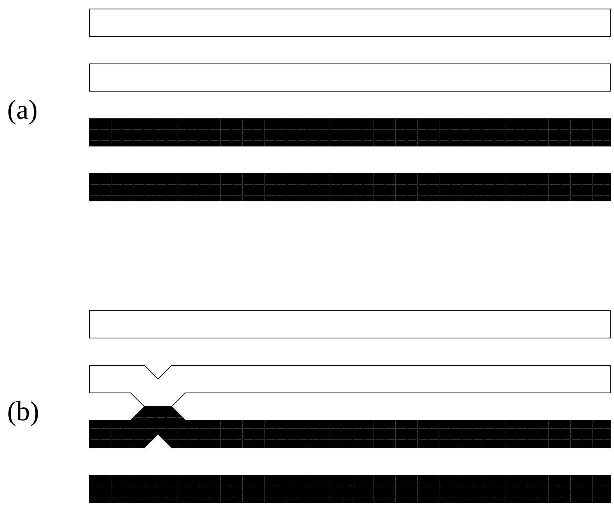

Figure 3.1: (a) Two copies of a chromosome pair. (b) Crossover between two non-sister chromosomes.

from which the $\mathcal{D}$ allele is inherited. Changes from 0 to 1 and vice versa occur with probabilities

$$\theta_i = \frac{1}{2}[1 - \exp(-2c_i/100)], \quad i = 1, \ldots, N.$$

This simulation method works because of the Markov property of the Poisson process. As an illustration, let $Y_{\mathcal{M}_i}$ denote the state at $\mathcal{M}_i$ and consider the chain (from one parent to one of the offspring) to the left of the disease locus. Then the transition probabilities will be of the form

$$\mathsf{P}(Y_{\mathcal{M}_K} = s_2) = \begin{cases} 1 - \theta_K & \text{if } s_2 = 0 \\ \theta_K & \text{if } s_2 = 1 \end{cases}$$

and for $i = 1, \ldots, K-1$

$$\mathsf{P}(Y_{\mathcal{M}_i} = s_2 \mid Y_{\mathcal{M}_{i+1}} = s_1) = \begin{cases} 1 - \theta_i & \text{if } s_1 = s_2 \\ \theta_i & \text{if } s_1 \neq s_2 \end{cases}.$$

The chain to the right of the disease locus will instead be specified by

$$\mathsf{P}(Y_{\mathcal{M}_{K+1}} = s_2) = \begin{cases} 1 - \theta_{K+1} & \text{if } s_2 = 0 \\ \theta_{K+1} & \text{if } s_2 = 1 \end{cases}$$

and for $i = K+1, \ldots, N-1$

$$\mathsf{P}(Y_{\mathcal{M}_{i+1}} = s_2 \mid Y_{\mathcal{M}_i} = s_1) = \begin{cases} 1 - \theta_{i+1} & \text{if } s_1 = s_2 \\ \theta_{i+1} & \text{if } s_1 \neq s_2 \end{cases}.$$

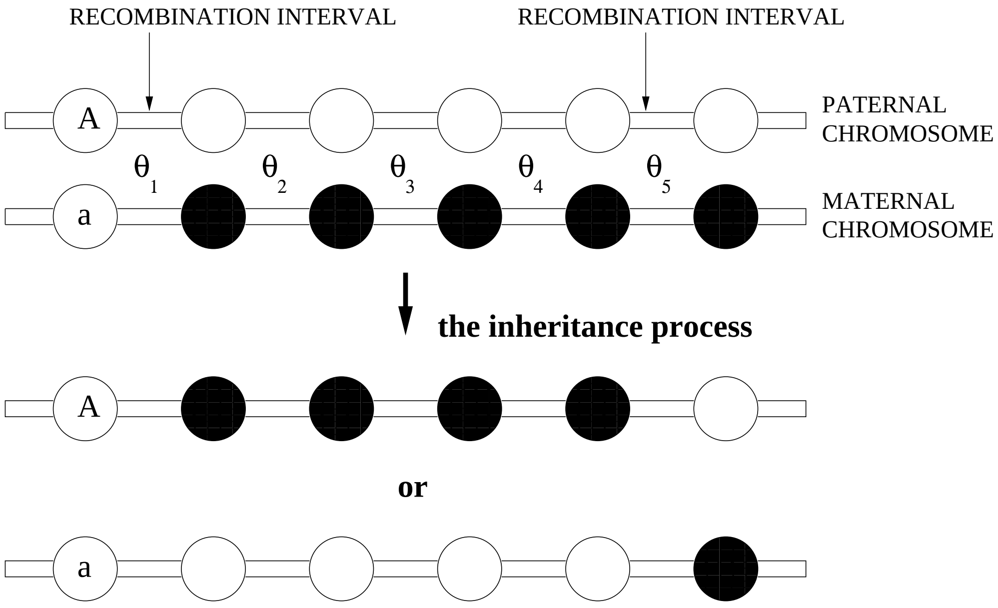

Figure 3.2: An example of the inheritance process from one parent to an offspring.

The Markov chains are independently simulated for the other offspring in the same manner.

Figure 3.2 shows an example of the inheritance process from one parent to an offspring. Here, the parent is heterozygous Aa at $\mathcal{D}$ with the trait allele on the paternal chromosome and the normal allele on the maternal one. Furthermore, let the alleles at the five marker loci $\mathcal{M}_{K+1}$,$\mathcal{M}_{K+2}$, $\ldots$,$\mathcal{M}_{K+5}$ to the right of $\mathcal{D}$ be $m_{K+1,1}$, $m_{K+2,1}$, $m_{K+3,1}$, $m_{K+4,1}$, $m_{K+5,1}$ on the paternal chromosome and $m_{K+1,2}$, $m_{K+2,2}$, $m_{K+3,2}$, $m_{K+4,2}$, $m_{K+5,2}$ on the maternal one. If two recombinations occur, one between $\mathcal{D}$ and $\mathcal{M}_{K+1}$, the other between $\mathcal{M}_{K+4}$ and $\mathcal{M}_{K+5}$, the offspring will inherit marker alleles $m_{K+1,2}$, $m_{K+2,2}$, $m_{K+3,2}$, $m_{K+4,2}$, $m_{K+5,1}$ if it has inherited the disease allele A at $\mathcal{D}$ and otherwise $m_{K+1,1}$, $m_{K+2,1}$, $m_{K+3,1}$, $m_{K+4,1}$, $m_{K+5,2}$.

Let us end this chapter with a brief summary:

- Sample parental genotypes G_M, G_F at $\mathcal{D}$, from the distribution

$$\mathsf{P}(G_M, G_F \,|\, ASP)$$

- Sample offspring genotypes O_1, O_2 at $\mathcal{D}$, from

$$\mathsf{P}(G_{O_1}, G_{O_2} \,|\, G_M, G_F, ASP)$$

- Assign parental genotypes at marker loci

$$\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_N$$

  according to population frequencies $\{q_{m_{ij}}\}$

- Pass down marker alleles from parents to offspring, according to the Markov chains previously described (and keeping track of from which chromosomes the $\mathcal{D}$ alleles were inherited).

Chapter 4

The Power Calculations

Given a specified genetic model (with the structure described in Chapter 2) marker data can be generated using the method from the previous chapter. This means that for a set of families or pedigrees we can make a large number of marker data simulations and for each replicate calculate some appropriate statistic to study the probability of finding linkage. Here the focus will be on the NPL-score statistic, which is a nonparametric measure of linkage, calculated by the GENEHUNTER software [KDRDL96] [KL98].

4.1 The NPL-Score

Assume for an ASP family that at a certain marker locus the father has genotype m_1m_2 and the mother m_3m_4 (see Figure 4.1).

If both of the offspring have inherited the m_1 allele from the father, they share one allele IBD (*identical by descent*), meaning that the offspring's paternal alleles have the same ancestral origin (come from the same ancestral chromosome). Since one allele is inherited from each parent, the two sibs will have either 0, 1 or 2 alleles IBD at the particular marker locus. Under the null hypothesis of no linkage, i.e. the disease and marker locus are at such a far distance from each other that alleles from these loci are inherited independently of each other, each affected sib-pair shares 0, 1 or 2 alleles IBD in proportions $\frac{1}{4}$, $\frac{1}{2}$ and $\frac{1}{4}$ ($Bin(2, \frac{1}{2})$). Under the alternative hypothesis (linkage) the rate of IBD sharing is expected to be higher (2 alleles IBD with probability greater than $\frac{1}{4}$).

The NPL-score statistic used in GENEHUNTER for testing linkage to a disease locus is based on this theory of IBD sharing and is for a fixed position

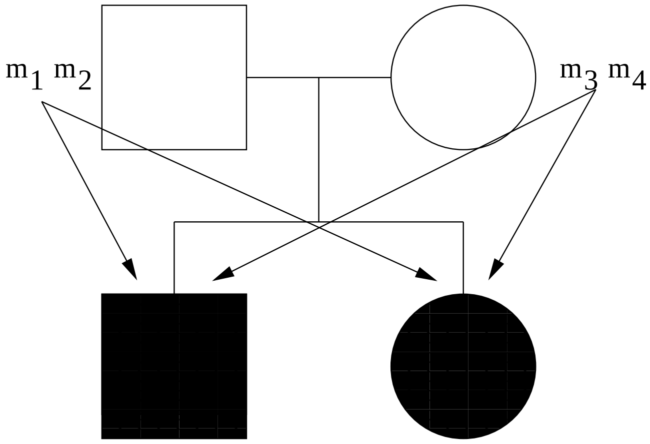

Figure 4.1: ASP family.

on the genome of the form

$$X = \sum_{i=1}^{m} Z_i/\sqrt{m},$$

where the sum runs over the families (pedigrees) and Z_i is a normalised variable, depending on the number of IBD alleles shared by affected individuals [KDRDL96]. To keep things as general as possible, Z_i should be looked upon as a variable assuming "large" (positive) values when the IBD sharing is high and "small" (negative) values when the sharing is low. By keeping this general level, the results can be transferred to other measures of the same structures. Therefore, the details regarding the NPL-score are treated in Appendix A, where Z_i also is calculated for a hypothetical ASP family.

If the IBD sharing can be unambiguously determined, $X \overset{\text{approx}}{\sim} \mathcal{N}(0,1)$, under the null hypothesis (due to the Central Limit Theorem). Otherwise (the IBD sharing not known with certainty), it can be shown that the variance of X is less than or equal to one [KDRDL96], i.e. treating X as $\overset{H_0}{\sim} \mathcal{N}(0,1)$ will provide conservative results. This NPL statistic is calculated at the predetermined N marker loci on the genome and the focus will therefore rather be on the vector

$$\boldsymbol{X} = (X^{(1)}, X^{(2)}, \ldots, X^{(N)})',$$

where $X^{(i)}$ is the score at $\mathcal{M}_i$ and the score at two adjacent loci are highly correlated. A more general form for the scores is

$$X^{(i)} = \boldsymbol{w}'\boldsymbol{Z}^{(i)},$$

where

$$\boldsymbol{Z}^{(i)} = (Z_1^{(i)}, Z_2^{(i)}, \ldots, Z_m^{(i)})'$$

and

$$\boldsymbol{w} = (w_1, w_2, \ldots, w_m)'$$

satisfying the condition

$$\boldsymbol{w}'\boldsymbol{w} = 1,$$

allowing different weights for different family structures and disease patterns. However in this thesis, only the case of equal weights

$$\boldsymbol{w} = (\frac{1}{\sqrt{m}}, \frac{1}{\sqrt{m}}, \ldots, \frac{1}{\sqrt{m}})'$$

will be treated.

Figure 4.2 shows the result from one GENEHUNTER run of simulated data for 50 sib-pairs without any parents (recessive disease locus with parameters $p_A = 0.2$, $f_{AA} = 0.5$, $f_{Aa} = f_{aa} = 0$, 24 marker loci at distance 5cM from each other, each with five equally likely alleles and the disease locus centred between the 12th and 13th marker). As seen in the figure, the computation of NPL-scores is not limited to the N marker loci, but can also be carried out in the intervals between the loci. However since we, when looking at the chromosome containing the disease locus, believe the peak (maximum score) in almost every case to be at a marker locus and the score between marker loci $\mathcal{M}_i$ and $\mathcal{M}_{i+1}$ to be $\leq \max\{|X^{(i)}|, |X^{(i+1)}|\}$ (an assumption supported by simulations and the limited amount of knowledge in the course of events between the markers), the interest is restricted to the vector $\boldsymbol{X} = (X^{(1)}, X^{(2)}, \ldots, X^{(N)})'$. This restriction will simplify the multivariate normal approximations in Section 4.2.

Powers can thus be obtained by doing a large number of simulations of m families and then look at the proportion of these simulations with scores above, say, the recommended thresholds for *suggestive* and *significant* linkage [LK95] (for details, see below). This will be denoted the *exact approach*[1]. If the main interest is to get the probability of finding linkage at any of the selected marker loci, the maximum score in each simulation replicate is

[1]The reader should not be confused by the word exact, which only means that there has not been any distributional approximations along the calculations like in the method about to be presented. The exact approach is based on simulations and is by no means exact.

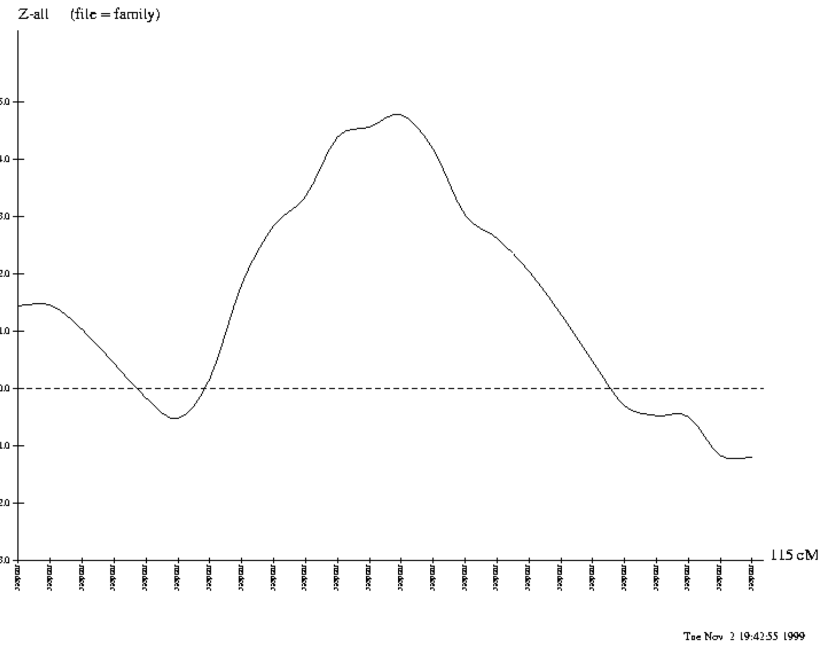

Figure 4.2: NPL-scores at 24 marker loci.

recorded. I.e. in each simulation generate N markers for m families, calculate NPL-scores at these markers and finally look at the maximum score only, at whichever marker it occurs, to see if it is higher than the recommended linkage thresholds. The power is approximated with the proportion of simulation replicates reaching the desired threshold, i.e. (given a threshold c),

$$\frac{\sum_{replicates} \mathbf{1}_{\{\max_{1\leq i\leq N} X^{(i)} \geq c\}}}{\#replicates},$$

where $\mathbf{1}$ is the indicator function defined by

$$\mathbf{1}_E = \begin{cases} 1 & \text{if } E \text{ is true} \\ 0 & \text{if } E \text{ is false} \end{cases}.$$

The threshold for significant linkage corresponds (roughly) to a multiple level of significance 0.05 when performing a whole (dense) genome scan, i.e. on average 0.05 false positives due to randomness on the whole genome [LK95]. The threshold for suggestive linkage gives a total expected number of one such false alarm for each genome scan. Using these thresholds for linkage, suggestive linkage will be found on average at one marker locus unlinked to the disease locus per genome scan and significant linkage on average at 0.05 unlinked loci. When analysing ASP families, the thresholds correspond to pointwise levels of significance of 2.2×10^{-5} (significant) and 7.4×10^{-4} (suggestive) [LK95], that is in the situation of interest the values 4.0854 and 3.1786 ($\Phi(1 - 2.2 \times 10^{-5})$ and $\Phi(1 - 7.4 \times 10^{-4})$ from the standard normal distribution).

4.2 The Multivariate Normal Approach

Although the marker simulations and the computations in GENEHUNTER can be performed within reasonable time, there still is a need for improvement. The reason being that a simulation method where the number of families m is fixed, forces the user to repeat the simulations using different values of m, until the desired power is obtained. Although the marker simulations only have to be performed for the largest m of interest, say $m_{\max}$, (by picking families from this set for all other $m \leq m_{\max}$) still the NPL calculations must be repeated for all m.

By sampling independent families under the alternative hypothesis H_1 corresponding to the specified model, a large number of Z-scores at the N locations on the genome (marker loci) can be obtained, i.e. we get a large number, say R, of observations from the vector

$$\boldsymbol{Z} = (Z^{(1)}, Z^{(2)}, \dots, Z^{(N)})',$$

where $Z^{(i)}$ denotes the score at $\mathcal{M}_i$, $i = 1, 2, \dots, N$. Let $Z^{(i,r)}$ denote the observed score at $\mathcal{M}_i$ from simulation r, $r = 1, 2, \dots, R$. For the score of one family the sample mean

$$\bar{\boldsymbol{Z}} = (\bar{Z}^{(1)}, \bar{Z}^{(2)}, \dots, \bar{Z}^{(N)})',$$

where

$$\bar{Z}^{(i)} = \frac{1}{R} \sum_{r=1}^{R} Z^{(i,r)}, \quad i = 1, 2, \dots, N,$$

and the sample covariance matrix

$$\boldsymbol{S} = \begin{bmatrix} S^{(11)} & S^{(12)} & \cdots & S^{(1N)} \\ S^{(21)} & S^{(22)} & \cdots & S^{(2N)} \\ \vdots & \vdots & \ddots & \vdots \\ S^{(N1)} & S^{(N2)} & \cdots & S^{(NN)} \end{bmatrix},$$

where

$$S^{(i_1,i_2)} = \frac{1}{R-1}\sum_{r=1}^{R}(Z^{(i_1,r)} - \bar{Z}^{(i_1)})(Z^{(i_2,r)} - \bar{Z}^{(i_2)}), \quad i_1, i_2 = 1, 2, \ldots, N,$$

can be used to get perfectly good estimates of its expectation

$$\boldsymbol{\mu} = (\mu^{(1)}, \mu^{(2)}, \ldots, \mu^{(N)})'$$

and covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma^{(11)} & \sigma^{(12)} & \cdots & \sigma^{(1N)} \\ \sigma^{(21)} & \sigma^{(22)} & \cdots & \sigma^{(2N)} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma^{(N1)} & \sigma^{(N2)} & \cdots & \sigma^{(NN)} \end{bmatrix}.$$

Being interested in m families, i.e. the vector

$$\boldsymbol{X} = (X^{(1)}, X^{(2)}, \ldots, X^{(N)})',$$

where

$$X^{(i)} = \sum_{j=1}^{m} Z_j^{(i)}/\sqrt{m}, \quad i = 1, 2, \ldots, N,$$

we use instead $\sqrt{m}\bar{\boldsymbol{Z}}$ to estimate the expectation of $\boldsymbol{X}$ and again $\boldsymbol{S}$ for the covariance matrix. This is motivated by the equalities

$$E[\boldsymbol{X}] = \frac{m\boldsymbol{\mu}}{\sqrt{m}} = \sqrt{m}\boldsymbol{\mu}$$

and

$$Cov(\boldsymbol{X}) = \frac{m\boldsymbol{\Sigma}}{(\sqrt{m})^2} = \boldsymbol{\Sigma}.$$

The next step is to use the central limit theorem to approximate the distribution of $\boldsymbol{X}$ by an N-dimensional multivariate normal distribution, i.e.

$$\boldsymbol{X} \overset{\text{approx}}{\sim} \mathcal{N}_N(\sqrt{m}\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

where the parameters are replaced by the observed estimates $\sqrt{m}\bar{\boldsymbol{Z}}$ and $\boldsymbol{S}$. This approximation can be used to get good power estimates. If for instance the main interest is the probability of finding linkage anywhere on the piece of the genome under study, we look at the distribution of

$$\max_{1 \le i \le N} X^{(i)}.$$

This is the problem of finding the distribution of the maximum in a N-variate normal distribution, which can be solved by sampling from such a distribution, or by numerical integration.

4.2.1 Sampling From A Multivariate Normal Distribution

Sampling from a $\mathcal{N}_N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution can be accomplished by several methods [BS72]. Below, one of them is described:

(1) Calculate $\boldsymbol{\Sigma}^{1/2}$. Any matrix $\boldsymbol{A}$, satisfying

$$\boldsymbol{A}\boldsymbol{A}' = \boldsymbol{\Sigma}$$

can be used as $\boldsymbol{\Sigma}^{1/2}$ and $\boldsymbol{A}$ can be obtained, using Cholesky factorization [BFS83]. To do so, $\boldsymbol{\Sigma}$ must be nonsingular. Problems will occur if any of the correlations will be equal to one[2].

(2) Generate a N-dimensional vector $\boldsymbol{V} = (V_1, V_2, \dots, V_N)'$ of independent random variables from the standard normal distribution ($V_i \sim \mathcal{N}(0,1)$, $i = 1, 2, \dots, N$).

(3) Form the vector $\boldsymbol{\Sigma}^{1/2}\boldsymbol{V} + \boldsymbol{\mu}$.

We have now obtained a random observation (N-dimensional) from the desired distribution $\mathcal{N}_N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. That (1)-(3) result in the right distribution is easily seen from the equalities

$$E[\boldsymbol{\Sigma}^{1/2}\boldsymbol{V} + \boldsymbol{\mu}] = \boldsymbol{\Sigma}^{1/2}(0, 0, \dots, 0)' + \boldsymbol{\mu} = \boldsymbol{\mu}$$

and

$$Cov(\boldsymbol{\Sigma}^{1/2}\boldsymbol{V} + \boldsymbol{\mu}) = \boldsymbol{\Sigma}^{1/2} Cov(\boldsymbol{V})\boldsymbol{\Sigma}^{1/2} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{I}_N\boldsymbol{\Sigma}^{1/2} = \boldsymbol{\Sigma},$$

where $\boldsymbol{I}_N$ is the $(N \times N)$ identity matrix,

$$\boldsymbol{I}_N = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}.$$

This means that

$$\boldsymbol{S}^{1/2}\boldsymbol{V} + \sqrt{m}\bar{\boldsymbol{Z}}$$

provides (approximately) a random sample of scores from the different locations on the genome. By getting a large number of such samples (by generating new $\boldsymbol{V}$'s only) and varying the number of families m we can receive nice power graphs depending on m. Observe that only a single set of simulations ($\boldsymbol{V}$ vectors) is needed and can be used for all m simultaneously, resulting in nice monotonicity properties.

[2] Will probably not happen since the sample covariance matrix is based on a large number of simulations.

4.3 Summary

In the previous sections, two power calculation techniques were introduced. First, the straight forward method called the the exact approach was described and by using an approximation step it was later improved into the *MVN approach* (multivariate normal approach). Before using them, we now summarize the two techniques.

Suppose one is interested in the probability of finding linkage at any of the marker loci $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_N$, for the different family sizes[3] $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_T$.

- The Exact Approach
  - For $\mathcal{S}_{t^*} = \max_{1\le t\le T} \mathcal{S}_t$ ASP families, simulate marker genotypes at the N marker loci, a large number R' of times.
  - For each of the R' replicates, calculate NPL-scores.
  - Approximate the power with the proportion of replicates yielding a maximum score higher than the given threshold c, i.e.

$$\frac{1}{R'} \sum_{r=1}^{R'} \mathbf{1}_{\{\max_{1\le i\le N} X^{(i,r)} \ge c\}},$$

    where $X^{(i,r)}$ is the score at marker i for replicate r.
  - For family sizes $\mathcal{S}_t < \mathcal{S}_{t^*}$, pick $\mathcal{S}_t \times R'$ families out of the $\mathcal{S}_{t^*} \times R'$ simulated ones and proceed as before.

- The MVN Approach
  - Simulate marker genotypes for a large number R of families.
  - Calculate NPL-scores for each family.
  - Approximate the mean vector and covariance matrix for the scores of one family by the sample mean vector $\bar{\boldsymbol{Z}}$ and sample covariance matrix $\boldsymbol{S}$. By using $\sqrt{\mathcal{S}_t}\bar{\boldsymbol{Z}}$ as the mean vector and again $\boldsymbol{S}$ as the covariance matrix, the approximate distribution for the scores of $\mathcal{S}_t$ families is obtained, using the multivariate normal distribution.
  - The power can be obtained by looking at the distribution of the maximum in a multivariate normal distribution. This is easily done by simulations. A single set of simulated observations from a standard normal distribution can be used to obtain the power for all family sizes.

[3] number of families

Certainly, a strategy to get simultaneous power curves, using a pure simulation technique is also possible, simply by adding families one by one to the sample and check for each sample size whether the relevant threshold is reached by any of the components. This gives a more complicated programme structure and moreover does not a priori give as smooth curves, as the multivariate normal approach.

When the number of family sizes to be tested grows, the MVN approach will be superior to the exact approach from the computation time perspective. However, the gain is not limited to computation time, since we also end up with a score distribution that may turn out useful. Next, the MVN approach is illustrated and tested in a simplified example.

Chapter 5

A Power Study for MS

MS (*Multiple Sclerosis*) is a disorder of the central nervous system, involving destruction of myelin. Reduced eyesight, difficulties with coordination and disturbance of balance are some of the recurring symptoms. Also the mental health is affected. The cause of the disease is believed to be genetic as well as environmental. MS is a worldwide disease, but with a varying prevalence. For instance, Scandinavia has a higher prevalence than, say, countries near the Equator. Studies have shown a change in occurrence risk of MS due to migration. The varying prevalences together with the influence of migration provide evidence for environmental factors involved in the disease. The fact that the occurrence risk for relatives of MS patients is much higher than for the population in general and that the risk for monozygotic twins is higher than for dizygotic twins, support the existence of genetic factors. In Scandinavia the population prevalence is about 0.1% and the sib prevalence about 4%. The age of onset is usually between 20-40 years. Although the course of the disease can be alleviated, there is today no existing cure for MS ([BO99], [McG97], [Nat94]).

5.1 The Family Study

It has been known for quite a long time that MS patients have what is called CSF-enriched oligoclonal lgG bands in higher proportions than healthy controls. Recent clinical studies have shown that healthy sibs of MS patients have ≥ 2 such bands in significantly higher proportions than unrelated controls. This condition (≥ 2 bands) will be denoted *MS-trait* [HAR+99].

In a recent study in Göteborg of 47 sib-pairs, each one involving a MS patient and one healthy sib, the MS-trait phenotype was observed in almost every MS patient (45 of 47 cases) and in 9 of 47 cases of the healthy sibs

[HAR+99]. The trait phenotype was observed in 2 of 50 healthy individuals (blood donors) leading to the suspicion of an MS-trait locus being one of many components giving rise to the MS disease. Since the bands are known to be caused also by some child diseases, it is hard to distinguish between the inherited and environmental factors behind the trait condition. However, the neurologists in Göteborg involved in this research have decided that it is worth testing the hypothesis that MS-trait is really an MS trait, caused at least partly by genetic conditions. A whole genome scan is planned and families with a MS patient as proband, a sib with MS-trait and if possible parents or extra sibs will be collected. In this example, we want to get an idea of the number of such ASP families that are necessary to collect, to have a fairly good chance to detect linkage if there is any, provided the rather optimistic scenario that this MS-trait is caused by a single dichotomic trait, which is either recessive or dominant and a penetrance event (in reality, it is believed to be much more complex and the power results in the sequel should be looked upon as optimistic upper bounds).

5.2 Power Calculations

The population prevalence K_P and the sib prevalence K_S for MS-trait, have been estimated by

$$\hat{K}_P = 0.04 \quad \text{and} \quad \hat{K}_S = 0.23,$$

yielding a λ_S about 5.8 (the estimate $\hat{K}_S$ is very uncertain).

As before, the disease allele is denoted by A. Two models have been used; one recessive and one dominant, each one with a λ_S around 6. In the recessive case, the parameters

$$p_A = 0.26, \quad f_{AA} = 0.6, \quad f_{Aa} = 0, \quad f_{aa} = 0,$$

are chosen. In the dominant case we let

$$p_A = 0.045, \quad f_{AA} = 0.45, \quad f_{Aa} = 0.45, \quad f_{aa} = 0.$$

In this example, $N = 10$ markers at 20cM distance from each other are simulated, each with 10 possible equally likely alleles. The disease locus is centred with 5 markers at each side ($K = 5$) and at the distance 10cM from the two closest markers, i.e. using the notation from Section 2, allele frequency vectors

$$\boldsymbol{q}_{\mathcal{M}_i} = (\frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}), \quad i = 1, 2, \dots, 10$$

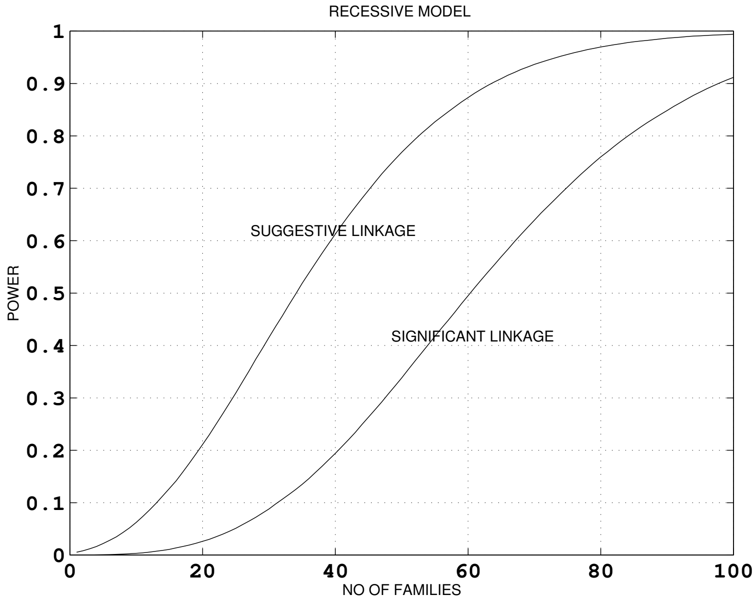

Figure 5.1: Power graph: The recessive case ($p_A = 0.26$, $f_{AA} = 0.6$, $f_{Aa} = f_{aa} = 0$). 2 affected offspring. 2, 1 or 0 parents (each with probability 1/3).

and the distance vector

$$\boldsymbol{C} = (20, 20, 20, 20, 10, 10, 20, 20, 20, 20).$$

In GENEHUNTER the 'all' scoring function is chosen (see [KDRDL96] and Appendix A for details) and NPL-scores are calculated at each marker locus.

To summarize: First marker data is simulated for 100000 ASP families, where each family includes two offspring (affected with MS-trait) and two, one or zero observable parents, each case with probability 1/3 (which seems like a reasonable assumption). Then, for each family NPL-scores are calculated at the 10 marker loci on the genome using GENEHUNTER. The sample mean $\bar{\boldsymbol{Z}}$ and sample covariance matrix $\boldsymbol{S}$ based on the 100000 simulations are then fairly good estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, respectively, where $\bar{\boldsymbol{Z}}$ is a 10-dimensional vector and $\boldsymbol{S}$ a 10×10 matrix. By sampling the distribution of $Z_{\max}$, according to the sampling scheme for a 10-dimensional multivariate normal distribution in Section 4.2.1 (100000 random samples from the multivariate normal distribution), the probabilities of reaching the thresholds for suggestive and significant linkage are estimated as graphs depending on the number of families (Figure 5.1 & 5.2).

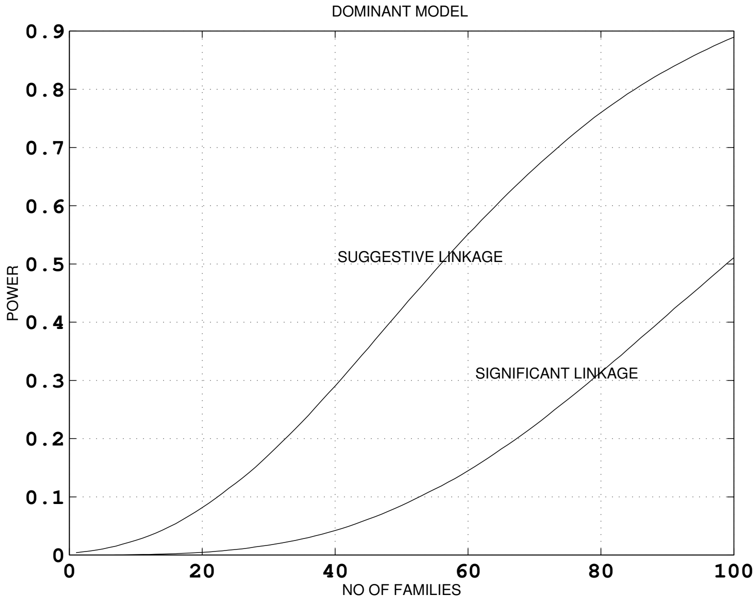

Figure 5.2: Power graph: The dominant case ($p_A = 0.045$, $f_{AA} = f_{Aa} = 0.45$, $f_{aa} = 0$). 2 affected offspring. 2, 1 or 0 parents (each with probability 1/3).

To find suggestive linkage with probabilities 0.5, 0.8 and 0.9 in the recessive case, about 35, 53 and 64 families will be needed. The corresponding number of families in the dominant case are 56, 86 and 103 (not shown in figure). To reach the critical value for significant linkage with probabilities 0.5, 0.8 and 0.9, about 61, 85 and 98 families are necessary in the recessive case. In the dominant case however, as many as 99 families will be needed just to reach this threshold with probability 0.5 (137 and 158 for the larger probabilities).

It may also be of interest on what locations the maximum scores are found. First, let us limit the interest to only concern the case of 60 families (using the MVN approach). The model did contain 10 marker loci in a row at 20cM distance from each other, and the disease locus centred between the 5th and the 6th marker locus. Thus, the disease locus is tightest linked to the marker loci 5 and 6, making us hope that a large part of the max scores are located at these loci. This turns out to be the case. Considering the recessive case we find, when plotting the locations of the first 1000 max scores, that there are some scores also located at the 4th and 7th marker locus and seven "outliers", two of them at marker 3, three at marker 8 and one each at marker 1 and 9 (Figure 5.3). If we instead focus on the first 1000 that reached the threshold for suggestive linkage, we get rid of four of these

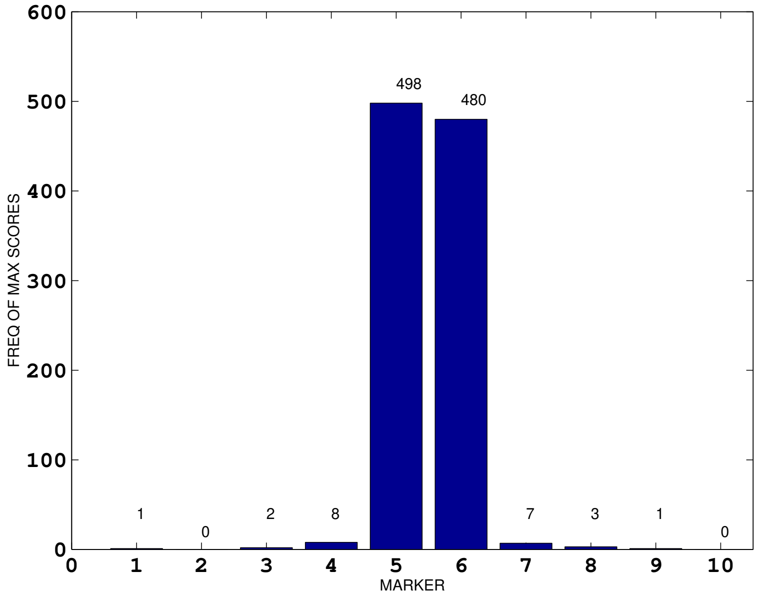

Figure 5.3: Locations of 1000 max scores (recessive case; $p_A = 0.26$, $f_{AA} = 0.6$, $f_{Aa} = f_{aa} = 0$).

outliers (Figure 5.4) and furthermore when considering those having reached significant linkage, there is only one max score (at marker 3) further than 30cM away from the disease locus.(Figure 5.5).

In the dominant case there is a larger variation of the max scores with values at all markers (Figure 5.6). Again, the variation decreases when considering only the ones passing the first threshold (Figure 5.7) and for the case of the second threshold there are, except for three (two at marker 3 and one at marker 9), no observations further away from the disease locus than 30cM (Figure 5.8).

Corresponding results for all number of families m are shown in Figure 5.9-12, where the proportion of the suggestive, the significant and of all the max scores that were found at marker 4-7 and 5-6, are plotted for the recessive and the dominant case, respectively.

Further simulations show that if a window is placed in front of locus 4 to locus 7, i.e. if the interest is restricted to the probability of finding linkage at the four marker loci within 30 cM from the disease locus, the loss in power in the recessive case will not be greater than 0.5% for any number of families (Figure 5.13). Furthermore, when only the powers obtained at marker loci 5 and 6 are considered, the loss is at most 1.5% (Figure 5.14).

The corresponding losses for the dominant case are about the same (Fig-

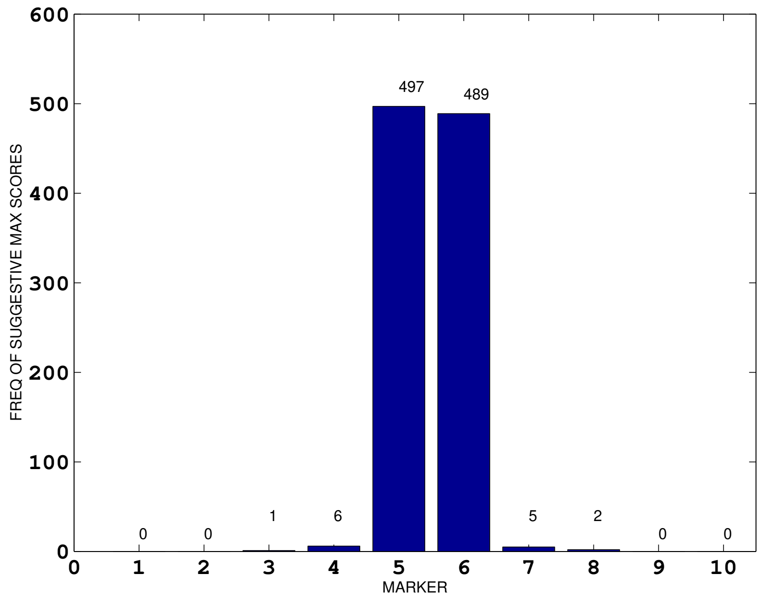

Figure 5.4: Locations of 1000 suggestive max scores (recessive case; $p_A = 0.26$, $f_{AA} = 0.6$, $f_{Aa} = f_{aa} = 0$).

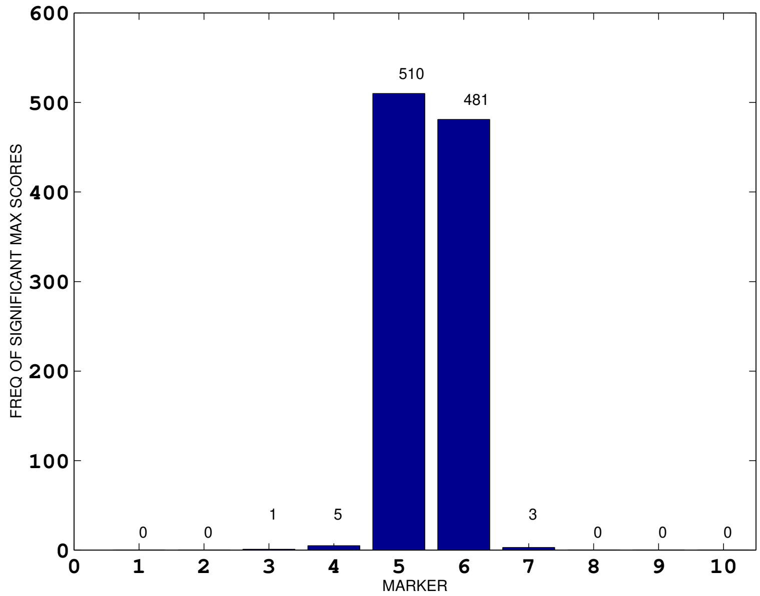

Figure 5.5: Locations of 1000 significant max scores (recessive case; $p_A = 0.26$, $f_{AA} = 0.6$, $f_{Aa} = f_{aa} = 0$).

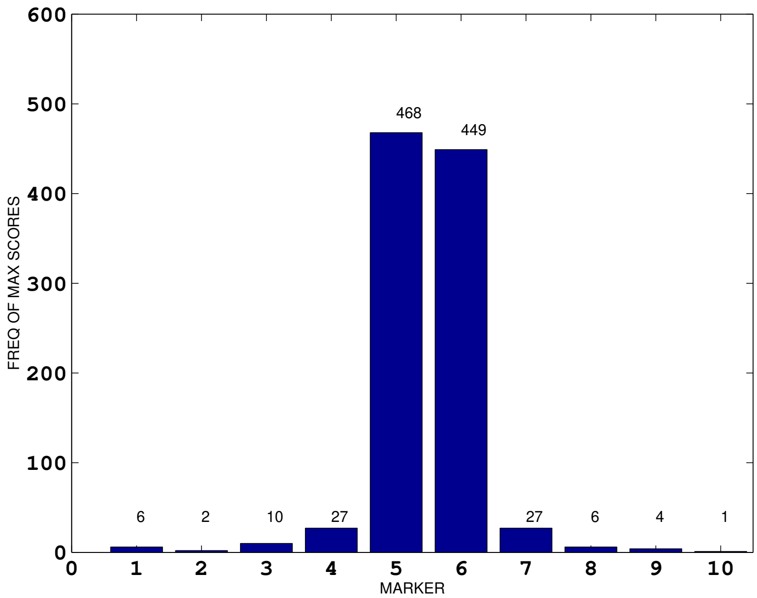

Figure 5.6: Locations of 1000 max scores (dominant case; $p_A = 0.045$, $f_{AA} = f_{Aa} = 0.45$, $f_{aa} = 0$).

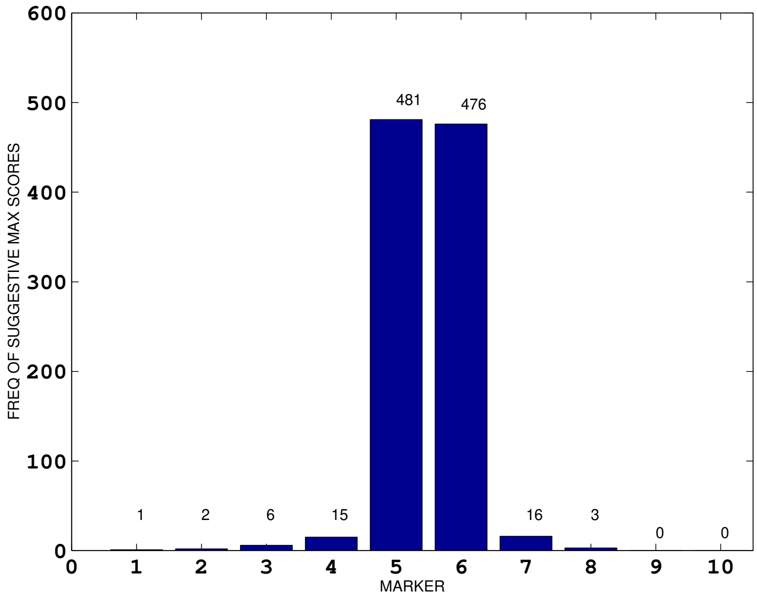

Figure 5.7: Locations of 1000 suggestive max scores (dominant case; $p_A = 0.045$, $f_{AA} = f_{Aa} = 0.45$, $f_{aa} = 0$).

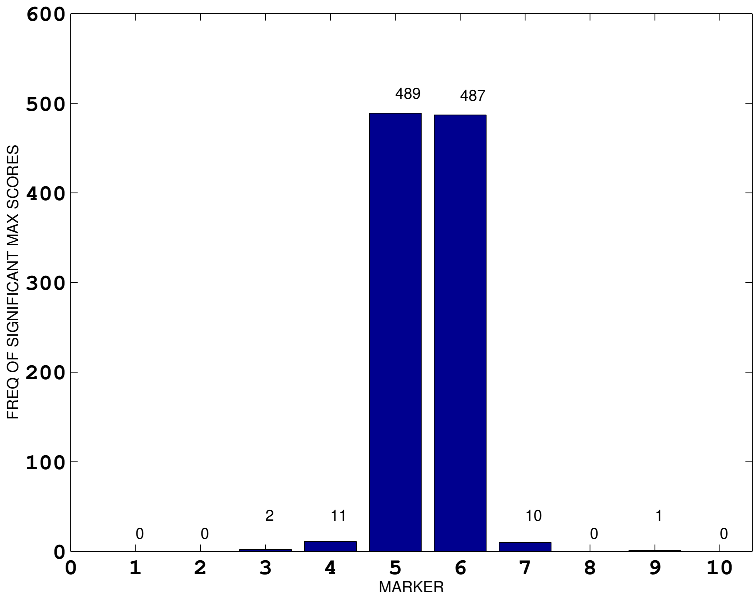

Figure 5.8: Locations of 1000 significant max scores (dominant case; $p_A = 0.045$, $f_{AA} = f_{Aa} = 0.45$, $f_{aa} = 0$).

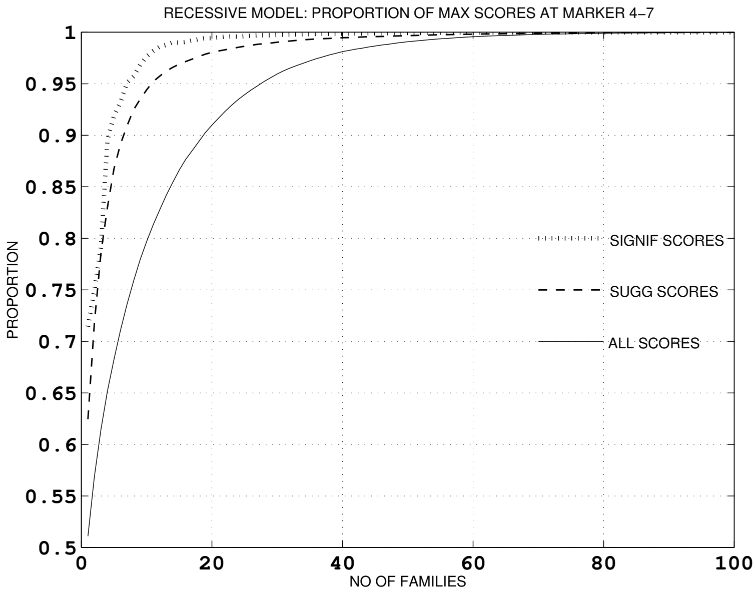

Figure 5.9: Max scores located at marker 4-7 (recessive case; $p_A = 0.26$, $f_{AA} = 0.6$, $f_{Aa} = f_{aa} = 0$).

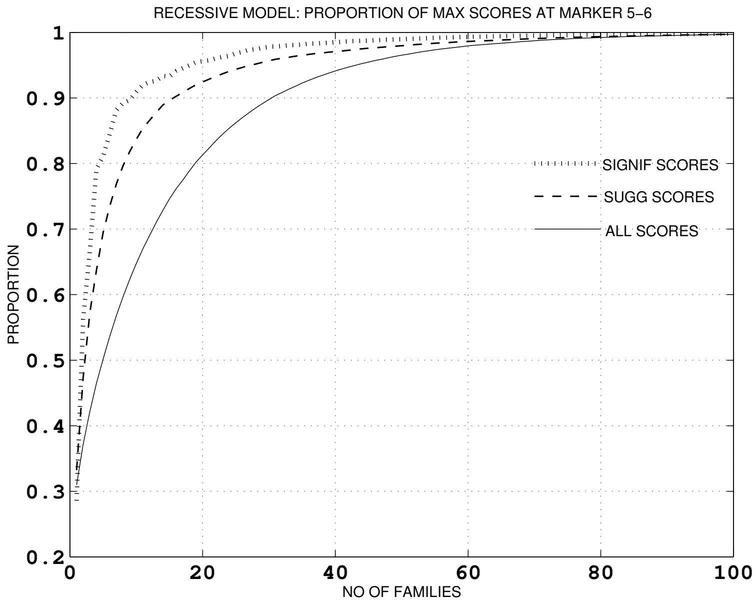

Figure 5.10: Max scores located at marker 5-6 (recessive case; $p_A = 0.26$, $f_{AA} = 0.6$, $f_{Aa} = f_{aa} = 0$).

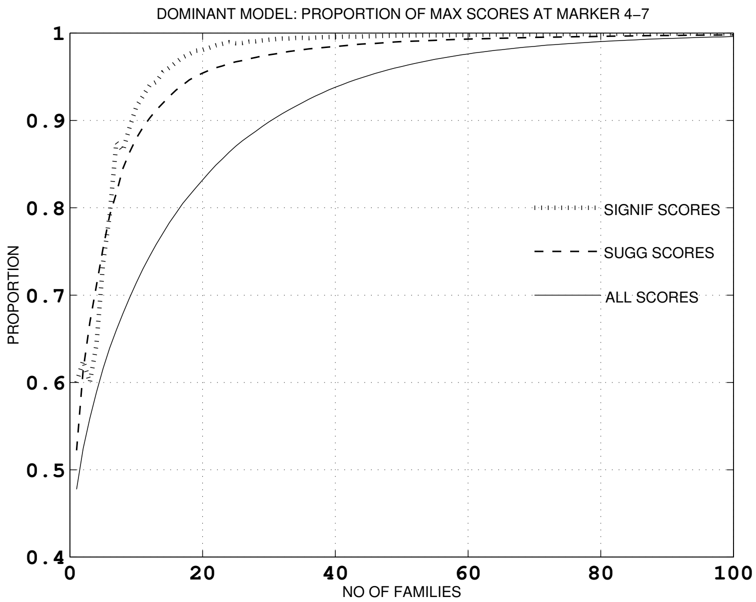

Figure 5.11: Max scores located at marker 4-7 (dominant case; $p_A = 0.045$, $f_{AA} = f_{Aa} = 0.45$, $f_{aa} = 0$).

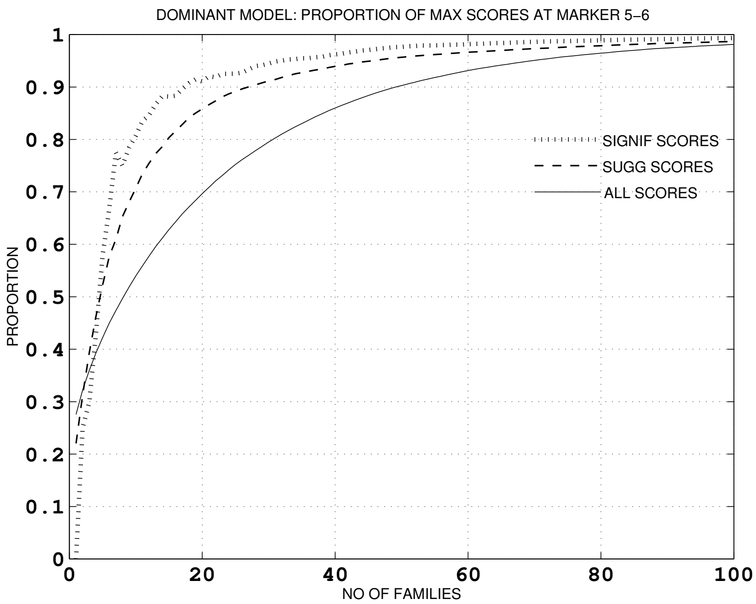

Figure 5.12: Max scores located at marker 5-6 (dominant case; $p_A = 0.045$, $f_{AA} = f_{Aa} = 0.45$, $f_{aa} = 0$).

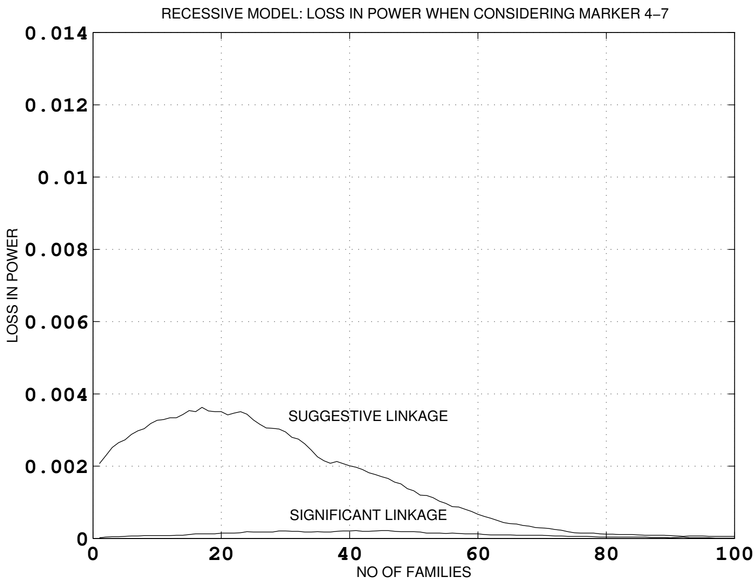

Figure 5.13: Loss in power when considering marker 4-7 (recessive case; $p_A = 0.26$, $f_{AA} = 0.6$, $f_{Aa} = f_{aa} = 0$).

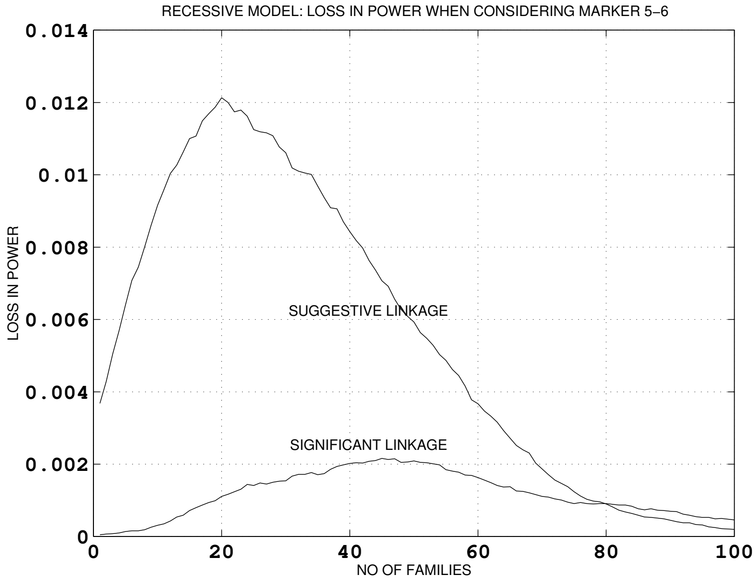

Figure 5.14: Loss in power when considering marker 5-6 (recessive case; $p_A = 0.26$, $f_{AA} = 0.6$, $f_{Aa} = f_{aa} = 0$).

ure 5.15 & 5.16).

5.3 Performance of the MVN Approach

How well does this MVN approach perform? To get a rough answer to this question, we make for both the recessive and the dominant case (using the models from the previous section), 10000 marker simulations with 20 families in each. For each simulation replicate we calculate NPL-scores and look at the proportions of suggestive and significant linkage obtained (this was denoted the exact approach). This is repeated for 40 and 60 families and comparisons between these values and the ones received in our graphs are printed in Table 5.1. Being based on a limited number of simulations, the exact approach is of course not “exact”. Hence the deviations between the two approaches are only natural, and would probably have been even smaller if the figures in both approaches had been based on a larger number of simulations. However, these deviations are quite small, stating that the use of the central limit theorem works out just fine. Since we use a finite number of simulations, there will also be some deviations between different runs of the MVN approach when different initial states for the random number generator are used. Figure 5.17 shows the power graph for the recessive case, previously plotted in Figure

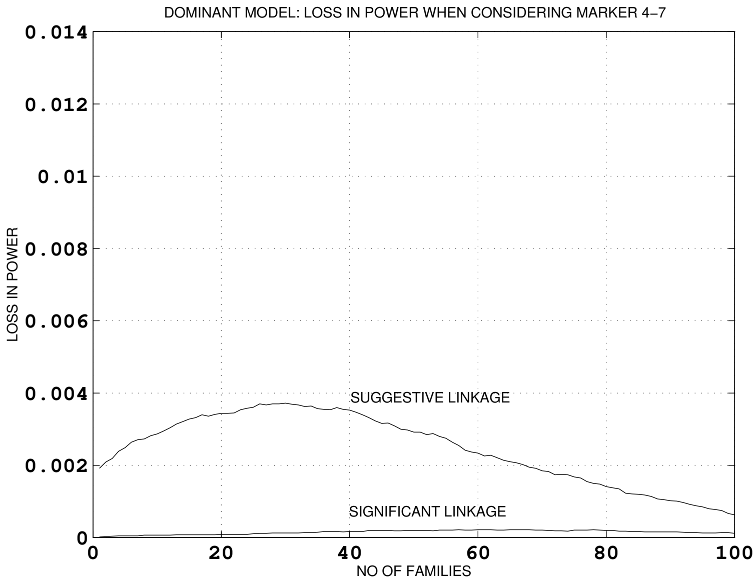

Figure 5.15: Loss in power when considering marker 4-7 (dominant case; $p_A = 0.045$, $f_{AA} = f_{Aa} = 0.45$, $f_{aa} = 0$).

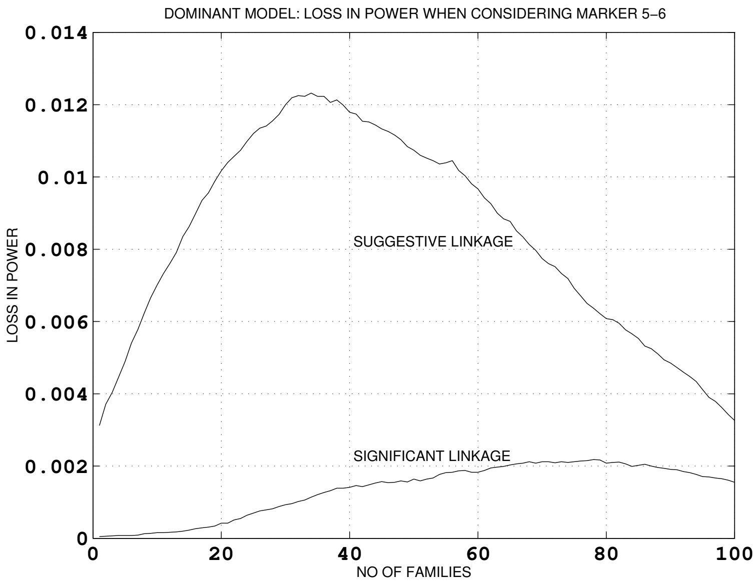

Figure 5.16: Loss in power when considering marker 5-6 (dominant case; $p_A = 0.045$, $f_{AA} = f_{Aa} = 0.45$, $f_{aa} = 0$).

case		MVN approach sugg link | signif link	exact approach sugg link | signif link
recessive	20 fam	0.2111 | 0.0262	0.2065 | 0.0164
	40 fam	0.6126 | 0.1941	0.6153 | 0.1868
	60 fam	0.8730 | 0.4949	0.8732 | 0.4918
dominant	20 fam	0.0817 | 0.0047	0.0725 | 0.0034
	40 fam	0.2903 | 0.0422	0.2843 | 0.0377
	60 fam	0.5513 | 0.1448	0.5499 | 0.1385

Table 5.1: Comparisons between the MVN and the exact approach. The figures displayed are the proportions of suggestive and significant scores, obtained from the MVN and the exact approach, respectively .

5.1, compared to a similar run based on a different set of random numbers.

Considering the purpose of the power study, i.e. to serve as guidance rather than to provide precise decimals for the power, the accuracy of the MVN approach seems sufficient.

5.4 Three Offspring

So far, we have only considered families with two offspring. How much better powers can be obtained when we instead use families with three offspring, where at least two of them are affected? The simulations in this section were performed by starting at the disease locus, randomly assign alleles to the parents, pass down alleles to the three offspring, check if the family satisfied some specified conditions and if that was the case, generate marker alleles according to the scheme in Chapter 3. Obviously, not the most elegant strategy (see Chapter 3), but justified by the circumstances under which the simulations were made. However, as in the two offspring case, no marker alleles are generated before a family has been accepted. A slight reduction in computation time was made by not assigning any alleles at $\mathcal{D}$ to the third offspring when none of the first two offspring were affected.

Considering the data from previous sections, with the slight change of only taking the four marker loci closest to the disease locus into account (see Figure 5.18). This reduction is made to get faster calculations and the loss in power is small enough to be acceptable (Section 5.2). Remember, the aim here is to compare the previously considered two offspring families to families with three offspring, thus the actual values are not that relevant.

If we in the recessive case compare the power graphs obtained from using

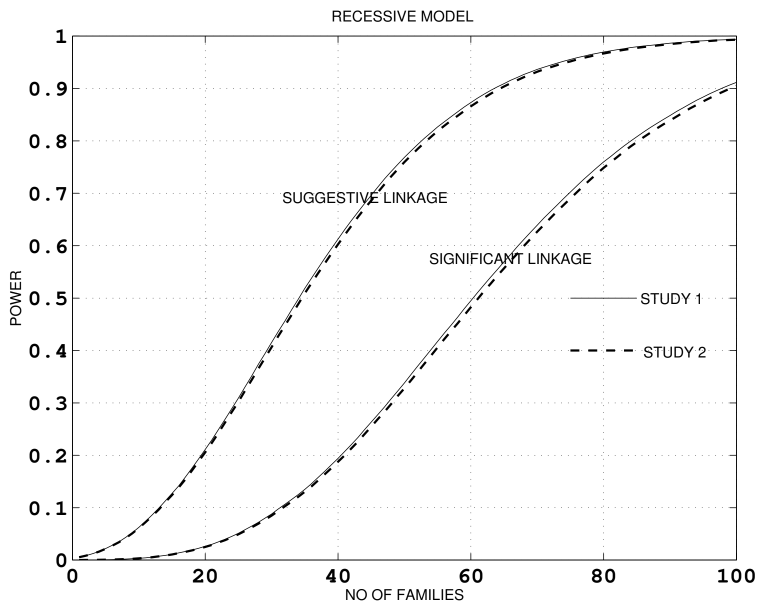

Figure 5.17: Two power simulations, using different sets of random numbers. Recessive case ($p_A = 0.26$, $f_{AA} = 0.6$, $f_{Aa} = f_{aa} = 0$). 2 affected offspring. 2, 1 or 0 parents (each with probability 1/3).

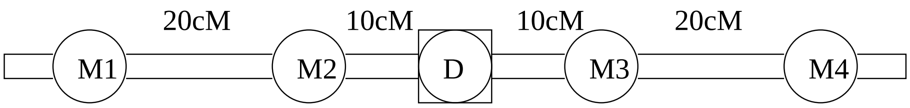

Figure 5.18: The disease- and the four marker loci considered in the comparisons between two- and three-offspring families.

two and three offspring, respectively (basing the normal approximations on 100000 simulated families and generate 100000 samples from these distributions), we notice that to obtain suggestive linkage with probability 0.8 we need 55 families with two or 39 with three offspring and with probability 0.9, about 66 families with two or 48 with three offspring (Figure 5.19). I.e. a gain of more than 15 families. To get significant linkage with probability 0.8 we need about 87 families with two or almost 63 with three offspring and with probability 0.9, 100 families with two or 73 with three offspring, i.e. a gain of more than 20 families (Figure 5.20).

In the dominant case, suggestive linkage is reached with probability 0.8 with 87 two offspring families or 68 with three offspring families, and with probability 0.9 with more than 100 or 81 families (Figure 5.21). To get the

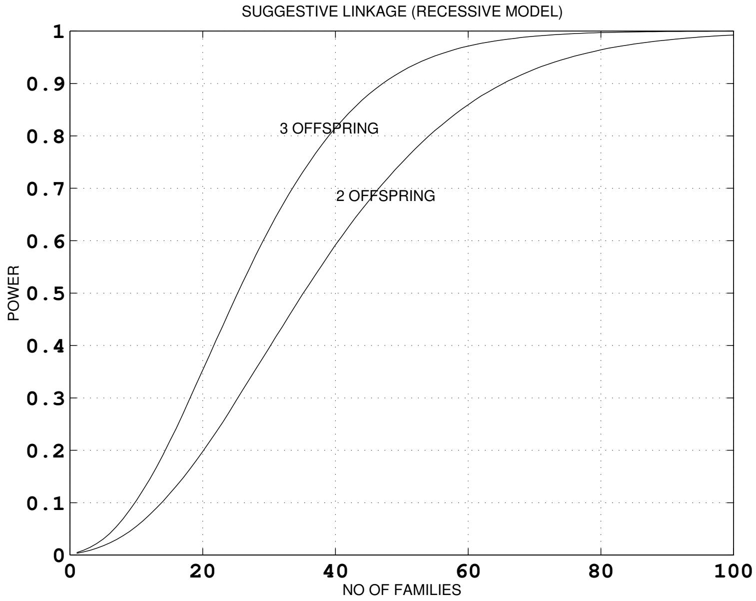

Figure 5.19: Power graph: The recessive case ($p_A = 0.26$, $f_{AA} = 0.6$, $f_{Aa} = f_{aa} = 0$). Families with 2 affected offspring compared to families with 3 offspring, where at least 2 are affected. 2, 1 or 0 parents (each with probability 1/3).

corresponding values for significant linkage we need larger family sizes than 100 (Figure 5.22).

The previous results showed how powers could be raised by taking advantage of a third offspring. However, one disadvantage of the NPL-score statistic is the fact that it does not make use of the IBD sharing information from a healthy offspring, i.e. bringing a healthy sib into the examination will only contribute to the information of the parental genotypes. The increase in powers seen before could in fact be the effect of a small number of families with three affected sibs. Discussing the gain from using an extra offspring, perhaps somewhat more interesting than the previous situation is to examine how much better result that can be obtained from using a healthy sib when parental data is missing. For ASP families without parental data, power comparisons are made between the cases with and without an extra healthy sib.

In the recessive case, 55 and 66 families consisting of only two offspring would be needed to receive suggestive linkage with probabilities 0.8 and 0.9. Using an extra healthy sib in each family, the corresponding sizes could be decreased to 38 and 45 (Figure 5.23). To receive significant linkage with

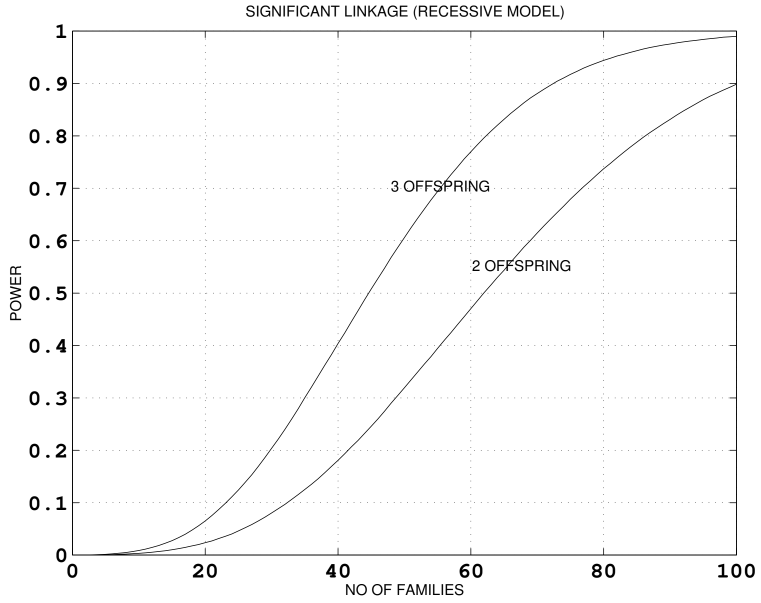

Figure 5.20: Power graph: The recessive case ($p_A = 0.26$, $f_{AA} = 0.6$, $f_{Aa} = f_{aa} = 0$). Families with 2 affected offspring compared to families with 3 offspring, where at least 2 are affected. 2, 1 or 0 parents (each with probability 1/3).

the same probabilities, 88 and about 100 families of two offspring would be needed, decreasing to 60 and 69 when adding the healthy sibs (Figure 5.24).

In the dominant case there is a need of about 73 affected sib-pairs to receive a suggestive linkage with probability 0.5, a figure to be compared to 58 if we add an extra healthy sib (Figure 5.25). To reach this level with probabilities 0.6, 0.7 and 0.8, the corresponding number of sib-pairs are 83, 94 and >100, but only 66, 76 and 87 with the added sib. Considering the threshold for significant linkage somewhat more than 100 sib-pairs would be needed just to reach it with probability 0.3 without the third (healthy) (Figure 5.26). Extracting the information from a third sib, significant linkage is reached with probability 0.3 and 0.4, using 81 and 91 families.

What conclusions could be drawn from these data? It is clear that powers can be increased when parental data is missing by using the information given from a third healthy sib, but this procedure will not reduce the set of individuals to be tested. When the parents were missing, the number of families to be tested to show suggestive linkage with probability 0.8 could be reduced from 55 to 38 in the recessive case, but the number of individuals increased instead from 110 to 114. To get suggestive linkage with probability

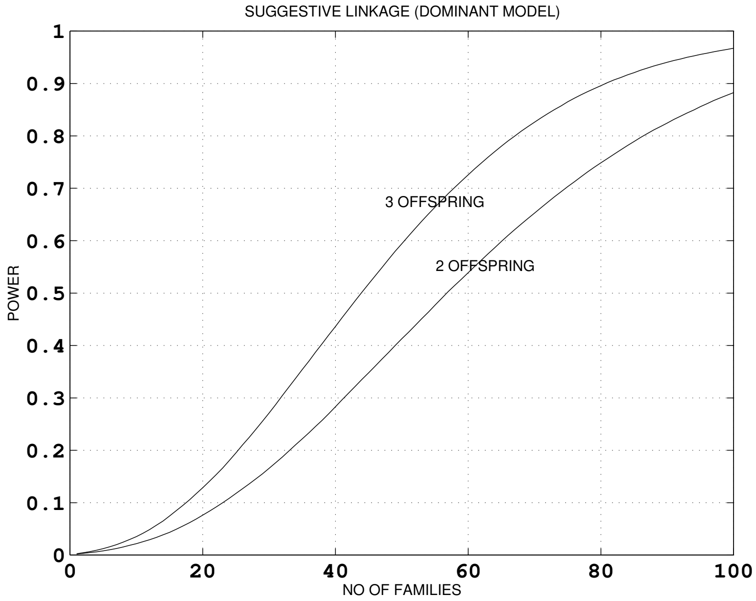

Figure 5.21: Power graph: The dominant case ($p_A = 0.045$, $f_{AA} = f_{Aa} = 0.45$, $f_{aa} = 0$). Families with 2 affected offspring compared to families with 3 offspring, where at least 2 are affected. 2, 1 or 0 parents (each with probability 1/3).

0.7 in the dominant case the number of families dropped from 94 to 76, but the total number of individuals increased from 188 to 228. Collecting another affected sib-pair would probably provide stronger evidence of linkage than the examination of two healthy sibs from two independent ASP families. Still, the data from healthy sibs can be the bridge between failure and success when the number of ASP families is limited. Another advantage is that this kind of data may be easy to get hold of. However, to extract all the available information a healthy sib provides, another measure or a modification of the NPL statistic would be needed. Remember that the only contribution from the healthy sibs is information of the parental genotypes, when those are not, or only partly, known. The fact that the sibs are healthy is not relevant and the extra sibs could therefore be ones with affection status unknown. Although not treated in this thesis, the reader should bear in mind the possibility of using different weights for different families (see [Nil99]).

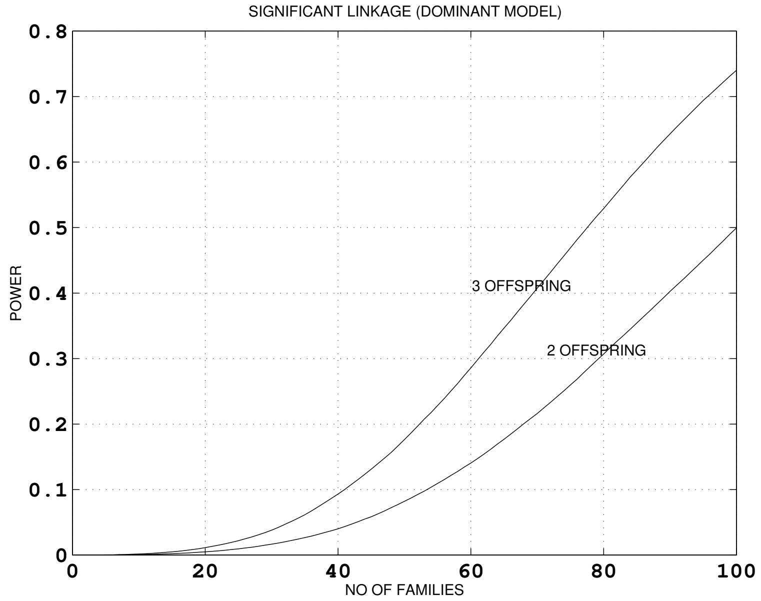

Figure 5.22: Power graph: The dominant case ($p_A = 0.045$, $f_{AA} = f_{Aa} = 0.45$, $f_{aa} = 0$). Families with 2 affected offspring compared to families with 3 offspring, where at least 2 are affected. 2, 1 or 0 parents (each with probability 1/3).

5.5 Some Comments

Although this example certainly is connected with reality, its main purpose in this thesis is to illustrate a method. To use an imaginary set of ten markers at 20cM distance between each neighbouring pair, is only a matter of convenience. In a genome scan, a tighter set of markers is usually chosen (10cM between two markers on average might be realistic). However, in the complete power study, different distances have been tested. The use of a total chromosome "length" of 180cM was somewhat arbitrary (the human chromosomes are about 140cM on average), but only a part of the chromosome, covering the disease locus, proved to be of value for the power calculations.

In the example, the parental data (when available) consisted of both marker genotypes and affection status. In many families, the parental data will consist only of marker genotypes, since a lumbal puncture must be performed to find out the affection status. Calculations (not shown here) show that this lack of data will in the dominant case reduce the power somewhat, but hardly influence the recessive case.

Since the IBD sharing and inheritance generally are not known with cer-

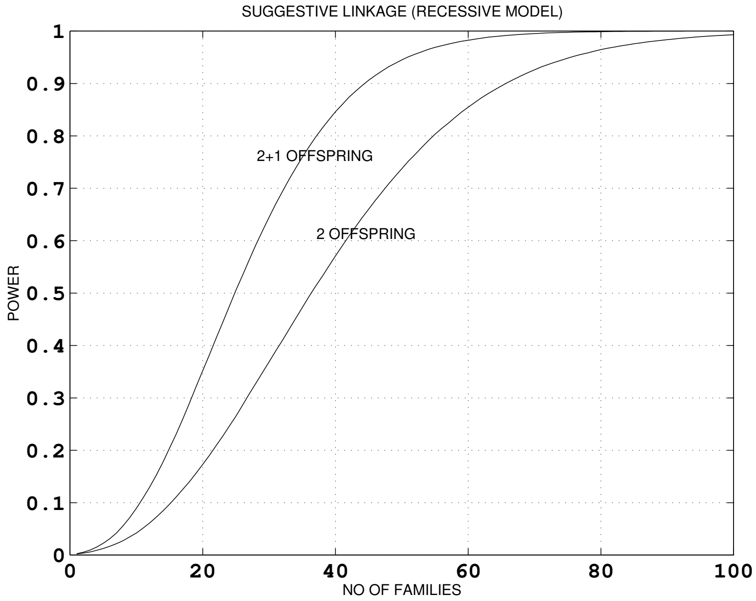

Figure 5.23: Power graph: The recessive case ($p_A = 0.26$, $f_{AA} = 0.6$, $f_{Aa} = f_{aa} = 0$). Families with 2 affected offspring compared to families with 2 affected + 1 unaffected offspring. No parents.

tainty, the marker allele frequencies will enter the NPL calculations. These frequencies are estimated through population data. If the frequencies vary between different populations, incorrect estimates might provide misleading powers. For instance, say that data for a large-scale study is collected in Sweden and Finland. Let m_{11} be an allele, existing in the Swedish population in the proportion $q^*_{m_{11}}$, compared to $q^{**}_{m_{11}}$ in the Finnish population. The estimate, based on the total population, will be a weighted average between $q^*_{m_{11}}$ and $q^{**}_{m_{11}}$. I.e. an incorrect estimate for all families, when $q^*_{m_{11}} \neq q^{**}_{m_{11}}$. What influence will this kind of falsities have on the powers?

5.5.1 Example

The parameter set, used in the simulations, included marker allele frequencies

$$\boldsymbol{q}_{\mathcal{M}_i} = (\frac{1}{10}, \frac{1}{10}, \dots, \frac{1}{10}), \quad i = 1, 2, \dots, 10.$$

What if the estimates

$$\hat{\boldsymbol{q}}_{\boldsymbol{\mathcal{M}}_i} = (\frac{3}{20}, \frac{3}{20}, \frac{3}{20}, \frac{3}{20}, \frac{3}{20}, \frac{1}{20}, \frac{1}{20}, \frac{1}{20}, \frac{1}{20}, \frac{1}{20})$$

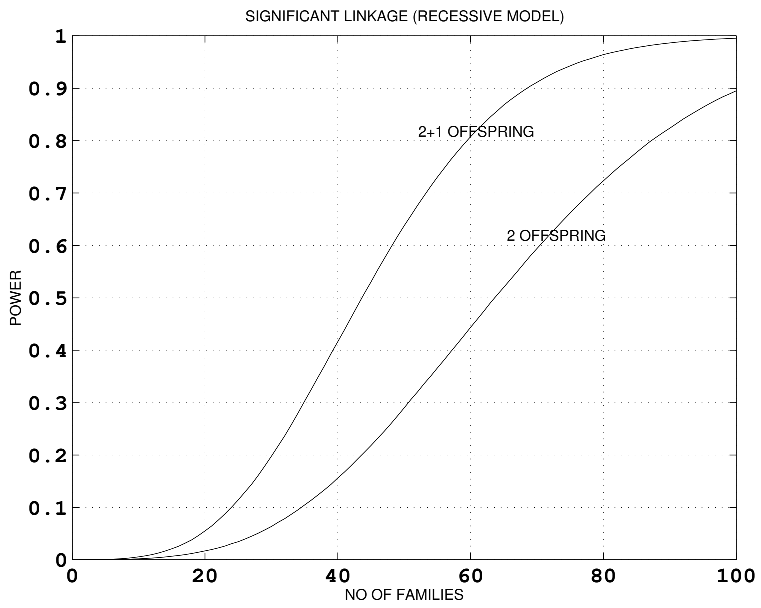

Figure 5.24: Power graph: The recessive case ($p_A = 0.26$, $f_{AA} = 0.6$, $f_{Aa} = f_{aa} = 0$). Families with 2 affected offspring compared to families with 2 affected + 1 unaffected offspring. No parents.

are used in the NPL calculations?

Figure 5.27 shows that these errors have a slight positive effect on the powers; in the recessive case almost negligible and in the dominant, less than 2.5% (the largest difference is reached in the dominant, significant case, for about 120 families). The error will probably also inflate the error of the first kind, so that the interpretation of suggestive and significant linkage will not be rigorously valid (see Chapter 6 below).

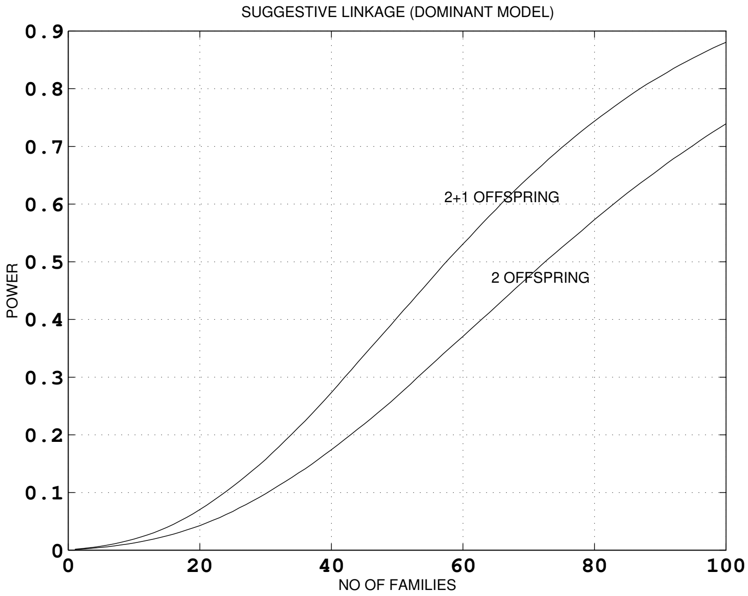

Figure 5.25: Power graph: The dominant case ($p_A = 0.045$, $f_{AA} = f_{Aa} = 0.45$, $f_{aa} = 0$). Families with 2 affected offspring compared to families with 2 affected + 1 unaffected offspring. No parents.

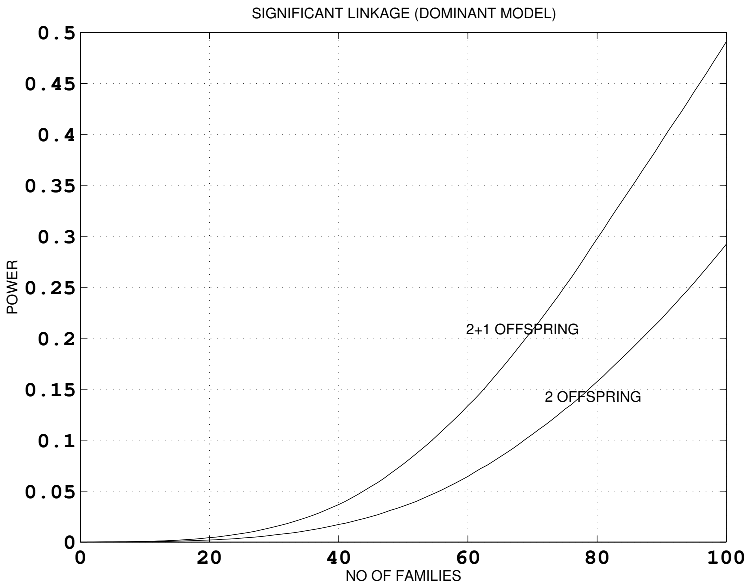

Figure 5.26: Power graph: The dominant case ($p_A = 0.045$, $f_{AA} = f_{Aa} = 0.45$, $f_{aa} = 0$). Families with 2 affected offspring compared to families with 2 affected + 1 unaffected offspring. No parents.

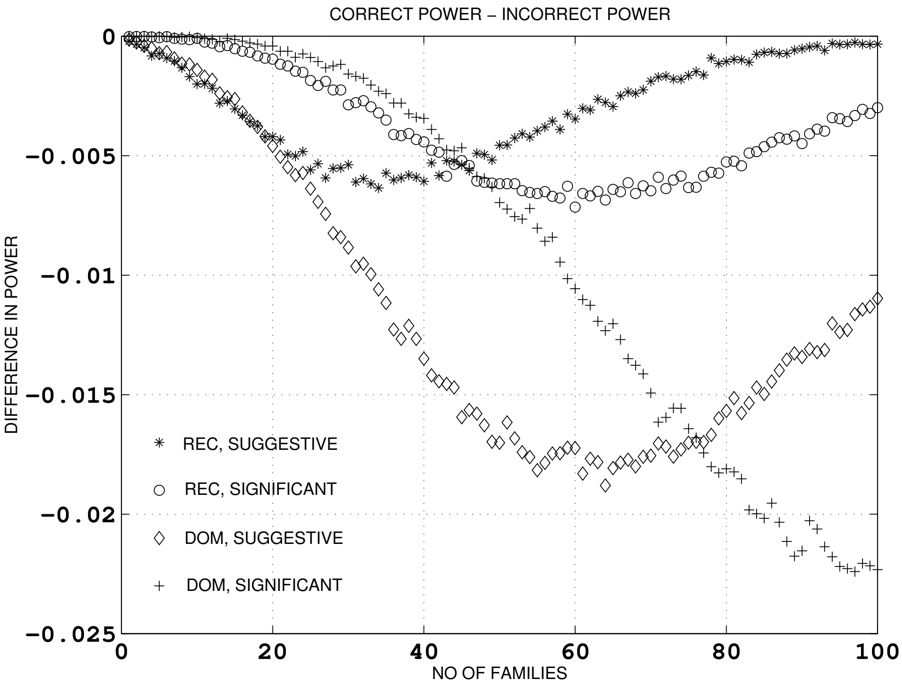

Figure 5.27: The difference between the correct power and the power obtained from using the incorrect marker allele frequencies (correct-incorrect).

Chapter 6

Further Applications

The thresholds previously described, called suggestive and significant linkage, are the proposed limits to use, to get the right multiple levels of significance when performing a whole genome scan [LK95]. Suppose the interest is restricted to a partial genome scan. For instance, knowledge about the disease (or previous scans) may point out a few chromosomes or a particular region on a certain chromosome that are of interest for the disease. If this in fact is the scenario, the limits based on whole scans seem unnecessarily strict, but of course some multiple adjustments must be made. The thresholds are depending on the number of chromosomes, the number of marker loci, their polymorphism and the distances (recombination fractions) between them. In other words, to choose the correct thresholds and pointwise levels of significance for the situation in question, is not a trivial problem. Of course, the Bonferroni method can always be used, i.e. using the pointwise levels $\frac{\alpha}{N}$ when dealing with N markers. This procedure will guarantee a multiple significance at most α (the probability of rejecting the null hypothesis at any marker when there is no linkage will be at most α, from Boole's inequality). However, due to the high dependence between adjacent marker loci, the multiple level will be much smaller than α and the inequality will not be very sharp. Instead the thresholds (and pointwise levels of significance) can be simulated, using the method described in previous sections, but now under the null hypothesis of no linkage (H_0).

Again, let $\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_N$ be the markers of interest and consider first the case when all the marker loci are located on the same chromosome. Remember that the random vector for the scores of one family was denoted

$$\boldsymbol{Z} = (Z^{(1)}, Z^{(2)}, \dots, Z^{(N)})',$$

where $Z^{(i)}$ was the score at $\mathcal{M}_i$, $i = 1, 2, \dots, N$. Under H_0, the mean vector is

$$\boldsymbol{\mu} = E[\boldsymbol{Z}] = (0, 0, \dots, 0)'$$

and the purpose of the simulations is restricted to estimate the covariance matrix

$$Cov(\boldsymbol{Z}) = \boldsymbol{\Sigma}.$$

As

$$E[\boldsymbol{X}] = \sqrt{m}\boldsymbol{\mu} = (0, 0, \dots, 0)',$$

(where

$$\boldsymbol{X} = (X^{(1)}, X^{(2)}, \dots, X^{(N)})'),$$

is the mean vector of scores for m families, and $\boldsymbol{\Sigma}$ is the covariance matrix for all m, we make (for m "sufficiently" large) the approximation

$$\boldsymbol{X} \overset{\text{approx}}{\sim} \mathcal{N}_N(\boldsymbol{0}, \boldsymbol{\Sigma}).$$

$\boldsymbol{\Sigma}$ is approximated by the sample covariance matrix $\boldsymbol{S}$, obtained from the simulations, where (using the notation from section 4.2) the elements are of the form

$$S^{(i_1,i_2)} = \frac{1}{R}\sum_{r=1}^{R} Z^{(i_1,r)} Z^{(i_2,r)}, \quad i_1, i_2 = 1, 2, \dots, N,$$

due to the knowledge of the mean vector. From this approximation, the threshold c_α, corresponding to a multiple level of significance α and satisfying

$$\mathsf{P}(\max_{1 \le i \le N} X^{(i)} \le c_\alpha) = 1 - \alpha,$$

can be obtained. The correct pointwise level will be

$$1 - \Phi(c_\alpha),$$

where Φ is the cumulative standard normal distribution function.

If instead H different chromosomes are considered, a "chromosomewise" level of significance (multiple level of significance for each chromosome)

$$1 - (1 - \alpha)^{1/H}$$

can be used. Due to the independence between marker loci at different chromosomes, the correct multiple level of significance

$$1-(1-(1-(1-\alpha)^{1/H}))^{H}=1-(1-\alpha)=\alpha$$

is obtained. Allowing for different levels of significance for different chromosomes, more general chromosomewise levels

$$\alpha_1,\alpha_2,\ldots,\alpha_H,$$

satisfying

$$\prod_{h=1}^{H}(1-\alpha_h)=1-\alpha,$$

can be chosen.

6.1 Example 1

Consider the following hypothetical example. Previous investigations concerning a certain disease have pointed out a particular region on a chromosome, covering about 50cM. To examine this region, 10 marker loci, each with ten equally likely alleles and the distance 5cM between each other, have been investigated. What pointwise levels of significance should be chosen to get a multiple level of 0.05?

Simulations of 100000 sib-pairs (no parents) yielded a sample covariance matrix

$$\boldsymbol{S}=\begin{bmatrix} s^{(11)} & s^{(12)} & \cdots & s^{(1,10)} \\ s^{(21)} & s^{(22)} & \cdots & s^{(2,10)} \\ \vdots & \vdots & \ddots & \vdots \\ s^{(10,1)} & s^{(10,2)} & \cdots & s^{(10,10)} \end{bmatrix}$$

and simulations (100000) of $\max_{1\le i\le 10} X^{(i)}$, where

$$\boldsymbol{X}\sim\mathcal{N}_{10}(\boldsymbol{0},\boldsymbol{S}),$$

suggested a threshold of 2.1681 to get the multiple level 0.05, the value obtained from plotting the *empirical distribution*

$$\hat{F}_{100000}(x)=\begin{cases} 0 & \text{if } x<o_{(1)} \\ \frac{i}{100000} & \text{if } o_{(i)}\le x<o_{(i+1)},\ i=1,2,\ldots,99999 \\ 1 & \text{if } x\ge o_{(100000)} \end{cases},$$

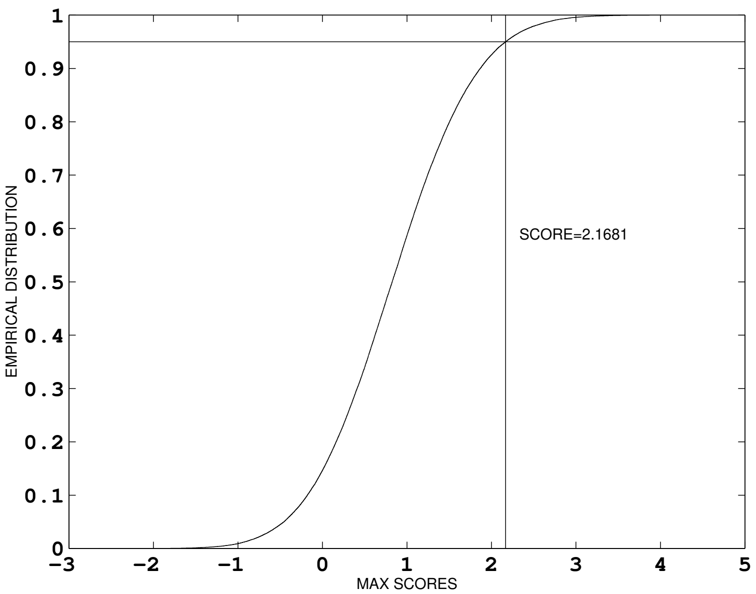

Figure 6.1: The empirical distribution of $max_{1 \leq i \leq 10} X^{(i)}$.

where $o_{(1)}, o_{(2)}, \dots, o_{(100000)}$ is the ordered sample from $\max_{1 \leq i \leq 10} X^{(i)}$ (see Figure 6.1). This threshold corresponds to a pointwise level of significance 0.0151, which can be compared to 0.005, using the Bonferroni method. The corresponding results when the number of possible alleles are varied between 5 and 10, and the distances between 5, 10 and 20cM (10 marker loci in each case), are listed in Table 6.1.

number of alleles	Distances (cM) 5	10	20
5	0.0221	0.0220	0.0237
10	0.0151	0.0135	0.0129

Table 6.1: Example 1: Pointwise levels of significance.

The pattern (or lack of pattern) in the table may seem illogical. Since the dependence between two marker loci decreases with the distance between them (high correlation between adjacent marker loci), the pointwise levels of significance should (in the case of a fixed number of marker loci) be larger the closer the marker loci are to each other. This is also the case when marker loci with 10 possible alleles have been used. However in the case of 5 alleles, the levels of significance are larger and the largest being obtained in the case of the largest distance. This "phenomena" can be explained by distributional differences. Remember that in the case of only one marker locus, the score is approximately standard normal distributed when the inheritance can be unambiguously determined, otherwise (inheritance not known with certainty) the variance is ≤ 1. Simulations (not included in the thesis) show that the variance increases with the polymorphism (increases with the number of possible alleles, at least if all alleles are equally likely), explaining the larger levels of significance in the case of 5 alleles (compared to 10 alleles). The lack of information can however be compensated by using a dense set of marker loci and analysing the inheritance at these loci simultaneously, pushing the thresholds for distances 5 and 10cM over the threshold for the case of 20cM when marker loci with 5 possible alleles are used (see Appendix A). Simulations show that the variances decrease when the distances increase. However, the differences are larger in the cases of 5 alleles than the ones involving 10 alleles.

6.2 Example 2

Consider the situation from Example 1, but instead of the assumption of equal allele frequencies, use

$$\boldsymbol{q}_{\mathcal{M}_i} = (0.25, 0.25, 0.20, 0.15, 0.15), \quad i = 1, 2, \ldots, 10,$$

in the case of five possible alleles and in the ten allele case, frequencies

$$\boldsymbol{q}_{\mathcal{M}_i} = (0.15, 0.15, 0.15, 0.15, 0.15, 0.05, 0.05, 0.05, 0.05, 0.05), \quad i = 1, 2, \ldots, 10.$$

The correct pointwise levels of significance, corresponding to a multiple level 0.05, are listed in Table 6.2. Again we observe a similar pattern.

number of	Distances (cM)		
alleles	5	10	20
5	0.0226	0.0225	0.0245
10	0.0163	0.0151	0.0148

Table 6.2: Example 2: Pointwise levels of significance.

Chapter 7

Concluding Remarks

Simulation based multivariate normal approximations, in connection with evaluation of the performance of nonparametric linkage analysis with the GENEHUNTER software, have been introduced. The approach has been tested in various examples and has been found a quite applicable tool, to use when studying a variety of different scenarios. It can be used to study many other situations than the ones described in this thesis and probably will be so in the future. The test statistic used to test for linkage throughout these pages was the NPL-score. However, too much importance should not be attached to the particular statistic. Similar approaches would probably be feasible when dealing with other statistics.

The possibilities to use the derived technique to study various robustness properties of the used statistics are also obvious. The effects of using the wrong genetic distances, the assumption of no sex dependence involved in the genetic distances and the Poisson crossover assumption, can be examined simply by changing the family simulations in an appropriate manner. In this context, the computational time aspects are important and the use of multivariate normal approximations thus motivated.

Bibliography

Bratley:1983

[BFS83] P. Bratley, B.L. Fox, and L.E. Schrage. *A Guide to Simulation.* Springer-Verlag, New York, 1983.

Bang-Oturai:1999

[BO99] A. Bang-Oturai. *Genetic Risk Factors in Multiple Sclerosis.* PhD thesis, Copenhagen University Hospital, 1999.

Barr:1972

[BS72] D.R. Barr and N.L. Slezak. A comparison of multivariate normal generators. *Comm ACM*, 15:1048–1049, December 1972.

Griffiths:1996

[GMS+96] A.J.F. Griffiths, J.H. Miller, D.T. Suzuki, R.C. Lewontin, and W.M. Gelbart. *An Introduction to Genetic Analysis.* W.H. Freeman and Company, New York, 1996.

Haldane:1919

[Hal19] J.B.S. Haldane. The combination of linkage values and the calculation of distances between the loci of linked factors. *J Genet*, 8:299–309, 1919.

Haghighi:1999

[HAR+99] S. Haghighi, O. Andersen, L. Rosengren, T. Bergström, J. Wahlström, and S. Nilsson. The "ms trait" phenotype. Submitted to the Journal of Neurology, 1999.

Kruglyak:1996

[KDRDL96] L. Kruglyak, M.J. Daly, M.P. Reeve-Daly, and E.S. Lander. Parametric and nonparametric linkage analysis: A unified multipoint approach. *Am J Hum Genet*, 58:1347–1363, 1996.

Kruglyak:1998

[KL98] L. Kruglyak and E.S. Lander. Faster multipoint linkage analysis using fourier transforms. *J Comput Biology*, 5:1–7, 1998.

Lander:1995

[LK95] E. Lander and L. Kruglyak. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nature Genetics*, 11:241–247, November 1995.

McGraw-Hill:1997

[McG97] *McGraw-Hill Encyclopedia of Science & Technology*. McGraw-Hill, 1997.

Nationalencyklopedin:1994

[Nat94] *Nationalencyklopedin*. Bra Böcker, 1994.

Nilsson:1999

[Nil99] S. Nilsson. *Two Contributions to Genetic Linkage Analysis*. PhD thesis, Chalmers University of Technology and Göteborg University, 1999.

Ott:1991

[Ott91] J. Ott. *Analysis of Human Genetic Linkage*. John Hopkins University Press, Baltimore, 1991.

Speed:1999

[SQ99] T.P. Speed and P.X. Quang. Statistics in genetics (week 2). Lecture notes, 1999.

Whittemore:1994

[WH94] A.S. Whittemore and J. Halpern. A class of tests for linkage using affected pedigree members. *Biometrics*, 50:118–127, March 1994.

Appendix A

The Nonparametric Linkage Score

On the following pages the NPL-score statistic(s) will be described. However, to keep the mathematics as simple as possible, only the situation of one (fixed) marker locus (see section A.3 for a comment on the multipoint case) is fully considered. The NPL-score was defined (at a certain marker locus) in the previous sections as

$$X = \sum_{i=1}^{m} Z_i/\sqrt{m},$$

where the sum runs over the m families (pedigrees) (the notations differ somewhat from the ones used by Kruglyak et al [KDRDL96]). For a fixed pedigree, let a be the number of affected individuals, g the number of non-founders, i.e. individuals with both their parents in the pedigree, and

$$\boldsymbol{v}_i = (x_{i1}, y_{i2}, \ldots, x_{ig}, y_{ig}), \quad i = 1, 2, \ldots, 2^{2g},$$

the possible inheritance vectors associated with the nonfounder individuals. Each $\boldsymbol{v}_i$ is a vector of 0s and 1s, where

$$x_{il} = \begin{cases} 0 & \text{if nonfounder } l \text{ has inherited the mother's maternal allele} \\ 1 & \text{if nonfounder } l \text{ has inherited the mother's paternal allele} \end{cases}$$

and

$$y_{il} = \begin{cases} 0 & \text{if nonfounder } l \text{ has inherited the father's maternal allele} \\ 1 & \text{if nonfounder } l \text{ has inherited the father's paternal allele} \end{cases}.$$

Kruglyak et al [KDRDL96] consider two different NPL-score statistics. NPL_{pairs} is based on IBD sharing between pairs of affected individuals, whereas NPL_{all}

takes IBD sharing between all the affected individuals in a pedigree into account in a more complicated manner, than just summing the pairs. Although NPL_{pairs} is the least complicated one of the two statistics, we have chosen to treat NPL_{all} first, since this is the statistic used throughout Chapter 5. NPL_{pairs} is described for comparisons and as a service to the reader.

A.1 NPL_{all}

For a fixed inheritance vector, let

$$\boldsymbol{u}_j = (u_{j1}, u_{j2}, \ldots, u_{ja}), \quad j = 1, 2, \ldots, 2^a$$

be the possible vectors where one allele is chosen from each affected individual and define $h(\boldsymbol{u}_j)$ as the number of nontrivial permutations $u_{j(1)}, u_{j(2)}, \ldots, u_{j(a)}$ of $u_{j1}, u_{j2}, \ldots, u_{ja}$, satisfying

$$u_{j(k)} \overset{\text{IBD}}{=} u_{jk}, \quad k = 1, 2, \ldots, a.$$

$h(\boldsymbol{u}_j)$ will assume large values when the vector $\boldsymbol{u}_j$ contains a "large" number of IBD alleles. If f is the number of founders, $h(\boldsymbol{u}_j)$ can also be expressed as

$$h(\boldsymbol{u}_j) = \prod_{l=1}^{2f} [\,(\#\text{times the } l\text{th founder allele appears in } \boldsymbol{u}_j)!\,] - 1.$$

For each inheritance vector $\boldsymbol{v}_i$,

$$S(\boldsymbol{v}_i) = 2^{-a} \sum_{j=1}^{2^a} h(\boldsymbol{u}_j)$$

is calculated [WH94]. $S(\boldsymbol{v}_i)$ will increase with the number of affected individuals sharing alleles IBD. Forming the weighted sum

$$\bar{S} = \sum_{i=1}^{2^{2g}} S(\boldsymbol{v}_i) \mathsf{P}_{data}(\boldsymbol{v}_i),$$

where $\mathsf{P}_{data}(\boldsymbol{v}_i)$ is the probability of $\boldsymbol{v}_i$ being the true inheritance vector conditioned on the marker data and under the null hypothesis of no linkage, Z is calculated as

$$Z = \frac{\bar{S} - \mu}{\sigma}.$$

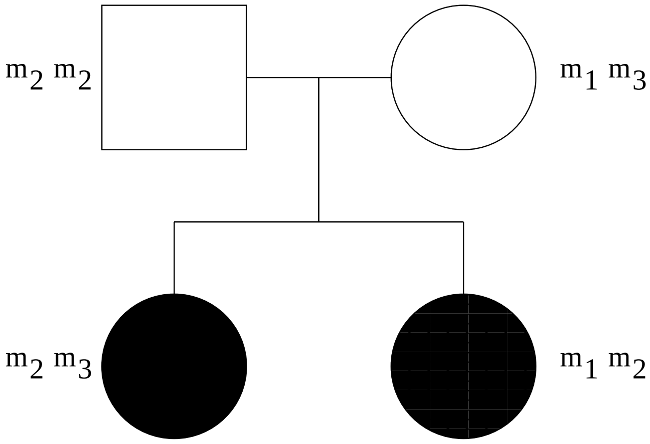

Figure A.1: An ASP family of two healthy parents and two affected daughters.

$\mu = E[\bar{S}]$ and $\sigma^2 = Var(\bar{S})$ are calculated under the hypothesis that all vectors $\boldsymbol{v}_i$ are equally likely, i.e. following the discrete uniform probability distribution $\mathsf{P}_{uniform}(\boldsymbol{v}_i) = 2^{-2g}$ (corresponding to the null hypothesis of no linkage to the disease locus in question). The total score X is referred to as NPL_{all}.

A.1.1 Example 1

Suppose a family consists of two healthy (unaffected) parents, the mother with genotype m_1m_3 and the father with genotype m_2m_2 at the marker locus (no information regarding which alleles are matern/patern) and two affected daughters with genotypes m_2m_3 and m_1m_2 (Figure A.1). The sibs share either zero or one allele IBD, since they obviously have inherited different alleles from the mother and we do not know what alleles have been inherited from the father. Using the notation previously specified, $a = g = 2$. Since $x_{i1} \neq x_{i2}$, $i = 1, 2$ (different alleles inherited from the mother),

$$\mathsf{P}_{data}((x_{i1}, y_{i1}, x_{i2}, y_{i2})) = 0, \quad \text{if } x_{i1} = x_{i2}.$$

The probability of the inheritance vector $(0, 0, 1, 0)$ is

$$\begin{aligned}\mathsf{P}_{data}((0,0,1,0)) &= \mathsf{P}((0,0,1,0) \mid data) \\ &= \frac{\mathsf{P}(data \mid (0,0,1,0))\mathsf{P}_{uniform}((0,0,1,0))}{\mathsf{P}(data)} \\ &= \frac{q_{m_2}^2 q_{m_1} q_{m_3} \frac{1}{16}}{\frac{1}{2} q_{m_2}^2 q_{m_1} q_{m_3}} = \frac{1}{8}\end{aligned}$$

and the remaining alternatives are equally likely, i.e.

$$\mathsf{P}_{data}((x_{i1}, y_{i1}, x_{i2}, y_{i2})) = \frac{1}{8}, \quad \text{if } x_{i1} \neq x_{i2}.$$

The relevant values are listed in Table A.1. Using these values,

i	$\boldsymbol{v}_i$	$\mathsf{P}_{uniform}(\boldsymbol{v}_i)$	$\mathsf{P}_{data}(\boldsymbol{v}_i)$	$s(\boldsymbol{v}_i)$
1	(0,0,0,0)	1/16	0	1/2
2	(0,0,0,1)	1/16	0	1/4
3	(0,0,1,0)	1/16	1/8	1/4
4	(0,0,1,1)	1/16	1/8	0
5	(0,1,0,0)	1/16	0	1/4
6	(0,1,0,1)	1/16	0	1/2
7	(0,1,1,0)	1/16	1/8	0
8	(0,1,1,1)	1/16	1/8	1/4
9	(1,0,0,0)	1/16	1/8	1/4
10	(1,0,0,1)	1/16	1/8	0
11	(1,0,1,0)	1/16	0	1/2
12	(1,0,1,1)	1/16	0	1/4
13	(1,1,0,0)	1/16	1/8	0
14	(1,1,0,1)	1/16	1/8	1/4
15	(1,1,1,0)	1/16	0	1/4
16	(1,1,1,1)	1/16	0	1/2

Table A.1: Calculated values for each inheritance vector.

$$\bar{s} = \sum_{i=1}^{16} s(\boldsymbol{v}_i)\mathsf{P}_{data}(\boldsymbol{v}_i) = \frac{1}{8},$$

$$\mu = E[\bar{S}] = \sum_{i=1}^{16} s(\boldsymbol{v}_i)\mathsf{P}_{uniform}(\boldsymbol{v}_i) = \frac{1}{4},$$

$$\sigma^2 = Var(\bar{S}) = \sum_{i=1}^{16} (s(\boldsymbol{v}_i) - \frac{1}{4})^2 \mathsf{P}_{uniform}(\boldsymbol{v}_i) = \frac{1}{32},$$

resulting in the score

$$z = \frac{\frac{1}{8} - \frac{1}{4}}{\sqrt{\frac{1}{32}}} = -\frac{1}{\sqrt{2}}.$$

The contribution of this hypothetical family is therefore a negative one.

A.2 NPL_{pairs}

The pair statistic is obtained by defining $S(\boldsymbol{v}_i)$ as the number of pairs of alleles from affected individuals that, given the inheritance vector $\boldsymbol{v}_i$, are IBD. Again, the score is obtained from calculating

$$\bar{S} = \sum_{i=1}^{2^{2g}} S(\boldsymbol{v}_i)\mathsf{P}_{data}(\boldsymbol{v}_i),$$

and

$$Z = \frac{\bar{S} - \mu}{\sigma}.$$

A.2.1 Example 1 (continued)

Using the pedigree from Example A.1.1., the new $s(\boldsymbol{v}_i)$-values are listed in Table A.2. From the new values,

i	$\boldsymbol{v}_i$	$\mathsf{P}_{uniform}(\boldsymbol{v}_i)$	$\mathsf{P}_{data}(\boldsymbol{v}_i)$	$s(\boldsymbol{v}_i)$
1	(0,0,0,0)	1/16	0	2
2	(0,0,0,1)	1/16	0	1
3	(0,0,1,0)	1/16	1/8	1
4	(0,0,1,1)	1/16	1/8	0
5	(0,1,0,0)	1/16	0	1
6	(0,1,0,1)	1/16	0	2
7	(0,1,1,0)	1/16	1/8	0
8	(0,1,1,1)	1/16	1/8	1
9	(1,0,0,0)	1/16	1/8	1
10	(1,0,0,1)	1/16	1/8	0
11	(1,0,1,0)	1/16	0	2
12	(1,0,1,1)	1/16	0	1
13	(1,1,0,0)	1/16	1/8	0
14	(1,1,0,1)	1/16	1/8	1
15	(1,1,1,0)	1/16	0	1
16	(1,1,1,1)	1/16	0	2

Table A.2: Calculated values for each inheritance vector.

$$\bar{s} = \sum_{i=1}^{16} s(\boldsymbol{v}_i)\mathsf{P}_{data}(\boldsymbol{v}_i) = \frac{1}{2},$$

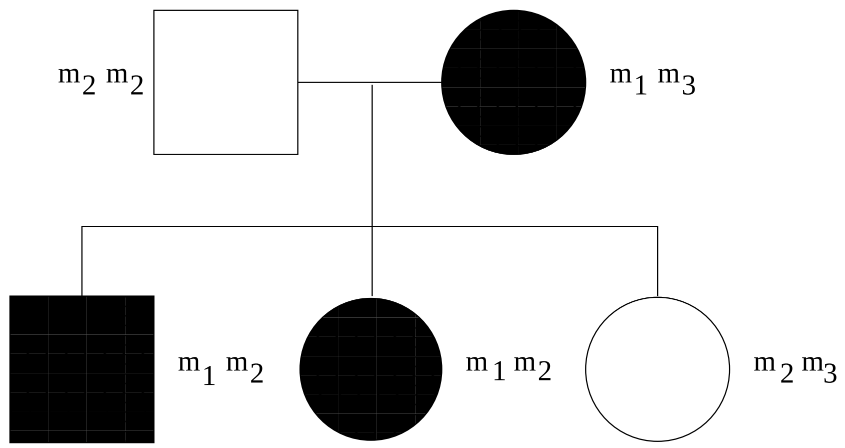

Figure A.2: An ASP family, consisting of two parents (the mother affected) and three offspring (two of them affected).

$$\mu = E[\bar{S}] = \sum_{i=1}^{16} s(\boldsymbol{v}_i)\mathsf{P}_{uniform}(\boldsymbol{v}_i) = 1,$$

$$\sigma^2 = Var(\bar{S}) = \sum_{i=1}^{16}(s(\boldsymbol{v}_i) - 1)^2\mathsf{P}_{uniform}(\boldsymbol{v}_i) = \frac{1}{2},$$

resulting in the score

$$z = \frac{\frac{1}{2} - 1}{\sqrt{\frac{1}{2}}} = -\frac{1}{\sqrt{2}},$$

i.e. $NPL_{all} = NPL_{pairs}$ in this example. This is always true for cases with only two affected individuals.

A.2.2 Example 2

A pedigree, which does not end up in the equality $NPL_{all} = NPL_{pairs}$, is seen in Figure A.2. Calculations show that NPL_{all} (≈ 0.894) gives a somewhat higher value than NPL_{pairs} (≈ 0.707). The contribution of this family is a positive one, regardless of which of the two statistics that are used.

If all the offspring were affected, we had instead received $NPL_{all} = -0.516$ and $NPL_{pairs} = -0.408$. Furthermore, if all the offspring had been affected and inherited the m_1 allele from the mother, the result had been $NPL_{all} = 1.549$ and $NPL_{pairs} = 1.225$ (calculations performed by GENEHUNTER).

A.3 Comments

The NPL-score Z at a certain marker locus for one family has mean 0. However, the variance will be 1 only in the case of full information, i.e. when the inheritance can be unambiguously determined, since μ $(= E[\bar{S}])$ and σ^2 $(= Var(\bar{S}))$ are calculated under full information. It can be shown that the variance of Z always is ≤ 1 [KDRDL96] (increasing with the amount of information). The use of the standard normal distribution thus provides conservative results. The problem can be reduced by using highly polymorphic markers.

The score functions were described only for the case of one marker. Analysing several marker loci, the probabilities $\mathsf{P}_{data}(\boldsymbol{v}_i)$ are calculated, taking all loci into account. This multipoint technique will also reduce the problems when not knowing the inheritance at each marker locus with certainty.

Appendix B

Comments on the Computer Implementations

The family simulation technique was implemented in MATLAB (version 5.3.0). However, the program is not yet fully developed and will be improved in the future. Just as an illustration, the version for generation of two-offspring ASP families takes as input

- Number of families to simulate
- Disease locus information $(p_A, f_{AA}, f_{Aa}, f_{aa})$
- Marker loci information $(N, \{n_i\}_{i=1,\dots,N}, \{\boldsymbol{q}_{\mathcal{M}_i}\}_{i=1,\dots,N})$
- Position of the disease locus (K)
- Distances in cM $(\boldsymbol{C})$
- Probabilities of including parents $(\mathsf{P}(0 \text{ parents}), \mathsf{P}(1 \text{ parent}), \mathsf{P}(2 \text{ parents}))$
- Parental affection status (known, unknown)

and returns as output a $(4 \times \text{No of families}) \times (6 + 2N)$ matrix where each row represents an individual and

- Column 1: family number
- Column 2: individual number
- Column 3: father's number if the row corresponds to an offspring, 0 otherwise

- Column 4: mother's number if the row corresponds to an offspring, 0 otherwise
- Column 5: sex (1 male, 2 female)
- Column 6: affection status (1 unaffected, 2 affected, 0 unknown)
- Column 7 to (6+2N): marker alleles

The output file is scanned into GENEHUNTER (version 1.3) [KDRDL96] [KL98], where NPL-scores are calculated. The last step, involving the multivariate normal calculations, we feel no more need to comment on (see section 4.2.1).